Reviewer comments are shown in black, *and author responses are shown in red. Line numbers in responses refer to the marked-up version of the revised manuscript.*

GMD Paper Review Response – Reviewer 1

Review 1 – Andrew Wickert

Skinner and coauthors' sensitivity analysis of landscape evolution models is a much needed addition to the geomorphic literature. Many evolution models often have been treated as sandboxes in which to experiment with quantified conceptual process understandings rather than as predictive models, with a few exceptions that include earlier work led by coauthor Coulthard using the CAESAR model. I also think that I have known of this work in progress for some time, as my lab donated some compute time at the request of coauthor Schwanghart.

Conceptually, I find the idea of a sensitivity analysis to be a very good approach. Often, landscape evolution models include competing processes designed to simulate the effects of real processes. Such additions typically include descriptions (qualitative and/or quantitative) if their impact on a landscape, but often do not include a more mathematical analysis of how these processes influence the solution space.

Thank you very much for the comments. We are glad you appreciate the complexity of the task – there is a great deal of information in and behind the paper and choosing what to highlight and discuss was a repeated issue in the paper's formulation.

The work of Skinner and coauthors is significant and publishable, but on reading their paper, I found multiple causes for concern. I could not find easy answers to the most major concerns, enumerated below.

• The sediment transport formula was the dominant source of uncertainty, but I believe that this may in part be because these sediment transport formulas were not appropriate for Tin Camp Creek (Australia).

– The grain size distributions for the rivers displayed significance of sand in the UK and a dominance of sand in Australia. Both the Wilcock and Crowe (2003) and the Einstein (1940's-50's) formulas are tested with coarse sand as the smallest grain size class. In the Australian case, about 50% of the sand is finer the grain size used to produce the sediment transport formula.

This is at the upper limit of the curve in Figure 6 of Wilcock and Crowe (2003), where their solution begins to bend more sharply but the data end. Therefore, there is great uncertainty and little constraint in the formulation.

We are not attempting to fit the model perfectly to each environment, and the use of Einstein and Wilcock and Crow (W&C) was largely because these are the formulae available in the un-modified model as downloaded. This acknowledged in lines 488-491 in the marked-up revised manuscripts –

"These were not selected as they represent the best fit for the catchments simulated but because they are the formulae available in the unmodified version of CAESAR-Lisflood."

There is a general point about sediment transport formulae and their applicability – that they generally perform well on the data they were generated from, but much less so when applied to other circumstances (eg Gomez and Church, 1989 and others since). So in taking any sediment transport rule out of its 'comfort zone' we will encounter issues.

For the second point, looking at Wilcock and Crowe (2003) they use sand fractions of 0.0005 to 0.64 m (Figures 1, 2 & 3, and P121) in their 48 flume experiments from which their formula was derived. For *our first catchment (the Swale) there is a pretty good agreement between the sediment grainsizes used and the W&C ones. In the second catchment – yes, there are finer sands and a greater proportion of sand in those simulations than used in the formula development. However, we find the same sensitivity to sediment transport formula in both basins, despite the different grain size mixes. Sediment transport formula choice has a much stronger impact than grainsizes as shown in our Figure 3.*

The issue of choice of sediment transport model and its applicability to any site is difficult to address as there is no general model. One can either calibrate to the site or develop a model specifically for the site which is impossible for most applications. This is no different to hydrological modelling where no model fits all situations. All model outputs are limited by this. We hope to get this point across in the paper.

The application to real basins – with their representative grainsizes is a nice thing to have – but the important finding is the similar sensitivity to sediment formula choice in two quite different settings/basins. Therefore, we argue that whilst our application of the W&C to the Australian example may be outside the limits of the initial W&C development – it does not mean that the overall finding is incorrect. There is certainly scope for a study/paper on the role of sediment transport rules in LEMs (see later comments).

– The Australian example has a dominance of sand. Are there bedforms that appear in the river? If so, could you discuss the role of their form drag, which to my knowledge is not included in your model, and how it could affect sensitivity to choice of sediment transport formulations?

To clarify, there are bedforms in the creek – and form drag is not included in the model. CAESAR-Lisflood will not generate realistic bedforms at this scale/resolution of application and these are not factored into the sediment transport formulas.

– I find your discussion of sediment transport in section 4.3 to be unnecessarily vague. It is not unusual to see in the landscape evolution modelling literature a statement to the effect of “sediment transport formulas are problematic and it is a difficult thing so the error is probably there”. Scientifically, this is unhelpful and in my opinion a little lazy. I think that here you have the opportunity to analyze why this is your major source of uncertainty, which is one way in which I hope this study can rise up above the others. Regardless of whether anyone trusts the form of your sediment transport formulations for the chosen grain size, form drag, etc., you have two mathematical formulations that must produce divergent outcomes for the set of provided hydrologic and topographic states. (I am presuming that over your 30-year time scales of interest, overall topography changes little.) Based on an analysis of these formulations, can you make a prediction of the factors that lead to this divergence?

These are really good points, well made. There is much scope (in another study) to look at how different sediment transport formulae respond (in the different settings) in this model that would be incredibly instructive to the LEM community and this is something we are working on (but first we need to look at how all the model parameters interact and influence model behaviour).

While not a direct comparison as prescribed in this paper but in the context of model testing Coulthard and Hancock have examined geomorphic change by comparing CAESAR and SIBERIA models over millennia. These tests show a lack of divergence and an equifinality of final form.

However, we would also like to emphasise that the primary purpose of this paper is a methods paper, introducing a method to assess the sensitivity of LEMs to changes in parameter values. Therefore, a detailed analysis of different sediment transport formulae is beyond the scope of this paper and also beyond the scope of GMD and would be better placed in an Earth surface based journal. However, we *agree that this would be incredibly valuable and a more in-depth test utilising more sediment transport formulae is planned. We see this paper as the introduction of a longer and larger set of experiments as part of the Landscape Evolution Model Sensitivity Investigation Project (LEMSIP) to which we would invite the community to contribute.*

• Your discussion notes that the sediment transport formula's importance may be overstated due to the smaller number of options for this than for the other variables. However, you do not analyze this possibility. or whether this would even lead to sediment transport formula remaining the dominant influence. Could you argue how your conclusion about sediment transport formulas is (or is not) still valid, considering this? I make a suggestion below (530-533).

This has been addressed in Section 4.5 – Limitations from Line 787.

"The main limitation of the MM is the subjectivity in selection of parameter values and ranges. Here, this has been mitigated by consistently selected ranges of +/- 50 % of a default value obtained from previous calibrations (where feasible). An issue emerges with categorical parameters, such as SED, where multiple values cannot be placed in spectrum across a range between minimum and maximum values. The MM has no formal method for dealing with such categorical parameters, so here it has been assumed that switching from one formula to another is a single iterative step change, and this would be the same even with more choices available. This reflects the purpose of the MM, which is to inform about the relative importance of choices of parameter values on the performance/behaviour of the model. However, to assess the impact of this single step-change assumption, we performed a further analysis, where it was assumed that switching formula was a change of four iterative steps. This analysis shows that the relative sensitivity of the model to the sediment transport formula choice becomes less important, with other parameters such as Manning's n Roughness and grain size sets increasing in relative influence (see Supplementary Material S2 for full results of this analysis)."

• Your premise is to test landscape evolution models, but the 30-year model run period is much shorter than most geomorphic models are used. Indeed, I wonder how much landscape evolution occurs, versus how much, over these time-scales, CAESAR can be thought of as a sediment-routing model with erosion or deposition being negligible (and therefore avoiding the nonlinearity in which changes in topography affect the long-term response of a LEM.) I think that this short time scale should be made explicit early in the paper. A discussion of how these results can (or cannot) be transferred to different time scales would be helpful as well.

Yes, this is a much shorter timescale than typically applied, but establishing a dominance on a process or set of processes over shorter times will still provide us with insight into longer term processes. Lessons learnt over the short time scales will apply over longer time scales. The 30 year time scale was largely chosen of model run time convenience – but it is also a timescale that is relevant to contemporary management decision making. Decision making processes based on geomorphic modelling require a far greater standard of understanding and handling of modelling uncertainties than currently employed by the LEM community, and lessons can be learnt from the hydrological and meteorological modelling communities.

In addition to these, the paper would be improved by a careful set of proofreading. It is repetitive in several places and includes a number of issues in both grammar and style. The overarching issues here are:

• Many proper nouns are capitalized; why?

• Your abbreviations should be used with an "s" to indicate whether or not they are plural in a given instance

• The Morris Method is mentioned 2–3 times before it is defined or described. Its description should be more closely tied to its in-text mentions.

This was also raised by reviewer 2 and we will make sure the method is defined earlier in the revised manuscript.

Morris Method is now defined much earlier, from Line 138.

• You define the difference between an objective function approach and a sensitivity analysis at multiple points; reduce this to one.

We have included a glossary of terms from Line 280.

• In general, many explanations are very “hand-wavey”. Please do a thorough read-through to reduce the fluff and improve the density of new information. If this is not done, it will be hard for a reader to see what interesting new conclusions you have come to.

We will sharpen up our descriptions.

Line-by-line and section-by-section comments are as follows:

Abstract: Why 3 paragraphs? I think you can shorten and tighten this.

We have rewritten the Abstract.

18. em dash after models; comma after example. Sensitivity Analyses one example here of something that is capitalized for reasons I don’t understand; I don’t think that this is just UK English.

Capitalised as it should be defining the acronym SA, but this is not included. Have edited accordingly.

47. I do not believe that your above references cover glacial or aeolian processes. Disregard if I am incorrect (does CAESAR include aeolian processes?); add references or remove these notes if I am. They are unimportant anyway to the landscapes that you are studying.

Have changed to just “Earth surface processes” as suggested by reviewer 2.

61. Comment: even few-parameter stream-power-based LEMs are quite heuristic. I do not think that your note here is unique to models with large numbers of parameters.

65. Correctness: an analysis cannot investigate, but you can.

79-85. I appreciate this list!

Thank you.

122. Define MM here or list section in which it is defined.

Now defined from Line 138.

128. Sentence fragment after comma

130. Incorrect in general: many landscape evolution models are not designed to be predictive over annual to decadal time scales. I find CAESAR to be quite unique in its time-scale flexibility, due (I believe) to its explicit integration of flow and sediment transport processes.

This has been edited (from Line 173) to "Moreover, some second-generation LEMs (e.g., CAESAR-Lisflood) simulate…" to reflect this.

131. I don't see what you mean by "multi-dimensional approach"

The phrasing is unclear. It means that the performance of the models needs to be assessed across all these timeframes. An LEM might produce reasonable behaviour when assessed at a millennial timeframe, but the behaviour at smaller timesteps could have no physical-basis. There could also be an element of equifinality where similar outputs from longer-term model simulations emerge after very different patterns of short-term behaviours. A model should be able to reproduce correct behaviours at all these timeframes and should be assessed on this basis. Text changed in Lines 173-175 to "necessitating data and methods to assess them across variable time scales."

133. When is an objective function not a score between observed and simulated values? Or do you mean that we can have synthetic observations?

Good point – have edited.

149-152. I think that point-based measurements must account for all of the complexity in the system, but may not be able to distinguish the source of the measured parameter's value.

A point-based measurement could never possible account for all the complexity in a system. For instance some changes may be spatially restricted – material is eroded, transported, and deposited wholly within the catchment so no signal of this change will ever cross the catchment outlet.

155. What is a width function?

According to Lashermes and Foufoula-Georgiou (2007), the "width function of a river network is a one-dimensional function which summarizes the two-dimensional branching structure of the river network. It represents the distribution of travel distances through the network and, under the assumption of constant flow velocity, the probability distribution of traveltimes. Thus its significance for understanding the hydrologic response of basins and the scaling characteristics of streamflow hydrographs is important". It is described in Hancock and Wilgoose (2001) as cited.

Lashermes, B., Foufoula-Georgiou, E., 2007. Area and width functions of river networks: new results on multifractal properties. Water Resources Research 43, W09405–W09405.

We did not feel this level of detail and description was required for the manuscript so have removed references to individual methods.

155. Cumulative area distribution – of what area?

Described in Hancock and Wilgoose (2001) as cited – "The cumulative area distribution (CAD) is a function defining the proportion of the catchment which has a drainage area greater than or equal to a specified drainage area. The CAD describes the spatial distribution of areas and drainage network aggregation properties within a catchment."

As above, we have removed all mentions to individual methods.

156. It is not possible to use variables as an objective function. One needs... a function. This may include these variables, of course!

These are not variables but values extracted from physical and numerical experiments, and compared as an objective function. This section has been rewritten (from line 224).

162. , and so are (comma)

162. "more objective" is vague: do you mean more in quantity or more as in better?

As above, this section has been rewritten.

165. "data" is plural. "data" is included twice as well, and the sentence is generally awkward.

167. Really? These data are not available? Not even in heavily-monitored experimental catchments? It is difficult to make sweeping statements, so I would ask you to prove this.

As above, this section has been rewritten to focus on the lack of methods rather than on data.

175. which –> that. This is an important distinction, and is often overlooked.

176. rm "will": tense confusion

185-186. "Medium" and "small" are nearly meaningless; could you provide catchment areas?

Have replaced with area values.

200. Is the rainfall time-series uniform in space or not? (I read later that it need not be, so please note this here, as this sets CAESAR ahead of other LEMs)

The rainfall time-series is derived from RADAR rainfall estimates and vary spatially for the Swale. The data has a spatial resolution of 5 km and a temporal resolution of 1 h (see 2.3.1.). Rainfall is uniform for Tin Camp Creek and based on one rainfall station (see 2.3.2.). This has been made clearer in the revised manuscript (Line 424-5 & 437).

204-205. Note how erosion and sediment transport are calculated here, as this is central to your conclusions.

We tried to limit our explanation of the model as its functionality is very well defined elsewhere and we have made no changes for this test. However, more detail on this aspect would be helpful and will include.

205. What is an "active layer system"?

Have expanded the description in the manuscript (lines 314-321). The active layer system allows for the storage of sub-surface sediment data by keeping it in strata going from a bed rock layer (if used), a base layer, several buried layers, and finally a top, active stream bed layer (Van de Wiel et al, 2007).

Van de Wiel MJ, Coulthard T, Macklin M, Lewin J. 2007. Embedding reach-scale fluvial dynamics within the CAESAR cellular automaton landscape evolution model. Geomorphology 90: 283–301. DOI: 10.1016/j.geomorph.2006.10.024

223-224. This is where singular vs. plural usage of acronyms can stand out.

240. What is the Design of Experiment? And how did it use R?

The Methods section describing the implementation of the MM has been rewritten, and this should be much clearer now (from Line 337).

250. is –> are

250. these constitute the Main Effect. (otherwise it is not clear how two things become one)

Have made it clearer that the mean of the elementary effects is the main effect, and the standard deviation of the elementary effects is separate. Line 365 – 369 – "After all 1600 tests have been performed, the main effect (ME) for each objective function and parameter is calculated from the mean of the relevant EEs – the higher the ME the greater the model's sensitivity. Alongside the ME, the standard deviation of the EEs is also calculated as this provides an indication of the non-linearity within the model."

266. elevation drop is also called "relief". I understand it here, but "relief" may be more common.

Have changed to "relief" (Line 419)

277. Grammar: Contrasting –> In contrast.

279. Comma after "Swale"

280. I thought that "rain gauge" is two words.

Changed.

282. Did you therefore use uniform precipitation for the Swale catchment?

No, rainfall series are based on NIMROD composite RADAR rainfall estimates and vary spatially. The resolution is provided in Section 2.3.1.

Table 1: (3 and 4) Do you mean to say that you prescribed a lateral erosion rate that is constant? This seems strange to me. (9) In/Out of what? (13) Evaporation or ET? (14) Roughness for channel, hillslope, or both?

CAESAR-Lisflood utilises a constant value for lateral erosion rate, which is different to the in-channel lateral erosion rate, and is part of the meander development section of the model. (9) the In/Out *difference is the ratio between water input and water output and is used to decide the timestep the model operates at. (13) Should be ET. (14) global value.*

311. It is definitively not qualitative, as it gives you numbers! Perhaps “quantitative but subjective”.

Good point, have changed this and similar references.

320. “Laws” is really strong. “Formulas” could be better. Or formulae, as you seem to be leaning Latin. Except with Germanic-leaning capitalization.

Yes, have changed to formulae throughout.

Figure 2: Perhaps note the D range of the data with which the sediment transport formulas were created

Have included a note about Figure 2 on Line 518-519 – “Note, that the grain size sets presented in Figure 2 contain non-cohesive silts and this requires an extrapolation of the sediment transport formulae (Van De Wiel et al., 2007).”

340-343. Section 1.3 doesn’t include what you state here. I also find it hard to believe that topography and discharge should have no relationship to geomorphic change. You will need to provide some evidence.

Should say “Section 1.2”, have edited.

343-344. This is no surprise: they studied equilibrium landforms, while you are studying 30-year time scales in which only extreme events can cause significant landscape change. In other words, your time scale removes the significance of topographic evolution and its associated feedbacks on the system.

Long-term landscape evolution is disproportionally influenced by successive extreme events. The short and long-term dynamics are intrinsically linked. Have removed this reference from the manuscript as it repeats parts of the Introduction.

348-350. Yes, you have mentioned this (note my general comment).

370-371. How have you assessed that 10 model years is a sufficient spin-up time?

It was just kept constant between the catchments. There is a wider issue here on the influence of spin-up time on model behaviour and uncertainties, another aspect we hope to look at as part of the LEMSIP work.

381, 383, etc. Consider giving full parameter names where possible to help the reader follow the text.

Originally we did, but the text became unwieldy as some names are long – we have kept them as acronyms in the revised manuscript but will take editorial advice on how these ought to be presented. Have included expanded forms within figure and table captions.

385. which –> that

Figure 3. Consider full model names; otherwise, you have converted numbers to codes, which readers will then have to cross-reference with your table.

Same as above – have included break-down of full names here in figure and table captions.

Figure 5. Is this catchment-wide elevation change, in-channel elevation change, or otherwise? In addition, are re you certain that mean elevation change is the appropriate metric? I can imagine that significant spatial variation in aggradation vs. incision could occur, and wonder how much this may affect your results. In addition, tens of cm of incision over 30 years seems very rapid to me: could you comment on this?

Catchment wide, sub-divided by the stream order definition – entire area under appropriate shadings.

We don't believe mean elevation change is necessarily a good metric but use it for illustrative purposes here to demonstrate that there are changes spatially on the 1600 DEMs generated by the simulations. Tens of cm of incision over 30 years is not uncommon in certain areas of both catchments – though changes in trunk streams in the Swale are controlled (in the field) by bedrock – which is deliberately not included in these simulations.

433. Small "s" on "LEMS"

443. How much do you trust gauged suspended sediment discharge and the associated rating curves? I do not know that these are so straightforward either. And if you mean bedload + suspended load, then I would argue that the data generally do not exist.

Indeed, have acknowledged this (line 688). Similar could also be said about the rainfall time-series used to drive the model.

4.1 (general). This section indicates to me that solving the LEM problem may be impractical due to the amount of time-lapse spatially distributed data required. Could you comment on this?

With present methods it might be – we have acknowledged this from Line 699 – "Some of the challenges of LEM output comparison are similar to those of meteorology/climatology and may require a shift in expectation from end users as to what is possible. For example, predicting detailed patterns of local erosion and deposition is akin to predicting weather (low comparability especially over longer time scales) but more general (spatial and temporal) patterns of basin change are similar to climate predictions (better comparability especially for longer time scales)."

But that does not mean we should not try and advance and seek new methods and techniques to try and address this. LEMs represent a potentially powerful tool for understanding geomorphic impacts due to changing climate, land use, and flood risk interventions, which could be applied for decision making purposes. The rewards for solving the LEM problem are worth pursuing a solution.

4.2. This section seems just to read, "we don't know how these models work and what the general rules are". It is OK to just write that! This seems to beat around the bush.

We have removed this section altogether.

456-457. Environmental models can be transferable between catchments. For example, I would argue that a thermal model is very transferable! Please be clear in what you mean by "environmental".

We are particularly referring to models of open systems which have variable parameters that are calibrated. The calibrated model cannot be directly transferred to another catchment, however similar *they are, and will need a new calibration. The same would apply to a sensitivity analysis. This section has been removed.*

473. Your sediment transport formulas do not include thresholds. Please explain how this compares to thresholded models if you include this point.

Wilcock and Crowe operates with a threshold – Einstein does not, but transport rates increase with a (loosely described) cubic function of stream power, in many ways mimicking a threshold in operation.

475. Your formulas include one that performed well in the Gomez and Church test and one that was not considered. What is the basis for using GC 1989, therefore, to declare that sediment transport formulas are not good?

Gomez and Church (1989) summarise that no formulae work well outside of the data upon which they were developed. We use this to illustrate a weakness of sediment transport laws that has been identified before.

477. Both of the formulas that you have employed are based on theory, and fundamentally on the force balance on a grain via the Shields number. I would suggest to not simply call these "empirical", but to actually note where the boundary between theory and empiricism lies. Indeed, these may be, for better or worse, some of the more theory-grounded components of a LEM! (Perhaps 2nd to the hydrodynamics)

Good points. We have edited this section from Line 728 – "The variation in the model performance can be explained by the derivation of the sediment transport formulae themselves, that are often theory-based but fitted to limited laboratory and field data, sometimes representing temporal averages over equilibrium conditions (Gomez and Church, 1989). The formulae do not, and were likely never intended to, represent the full variation of actual flow conditions in natural river. As LEMs commonly amalgamate a set of geomorphic models or transport formulae, their performance hinges in the a number of individual model components. Therefore, when applied to different situations, they may not be appropriate (Coulthard et al., 2007a)."

479. How do you know that they were not intended to represent variations in flow conditions? This statement is inconsistent with the fact that the underlying experiments have been performed at a wide range of tau=tau$_c$ ratios. You should be more specific or remove this comment.

Our wording is "full variation of flow conditions". It is true that the formulae are derived and tested over a wide range of flow conditions, but not the full range that might be experienced in reality. Have changed to "The formulae do not, and were likely never intended to, represent the full variation of actual flow conditions in natural river" from Line 731.

489-490. Non-stationarity in hydrologic models seems a bit off-topic here.

We do not believe this is off-topic here, but probably requires a lot more expansion and context. We have removed this reference.

482-493. Do you think that the issue of scaling and calibration should deserve at least its own paragraph, if not its own section?

Yes – have moved this to its own section (4.3) from Line 737.

507-514. Do you mean that LEMs should follow hydrologic models' approach to uncertainty estimation in general (there are many such approaches), or specifically Lisflood-LP, and why? In addition, this paragraph gives little information about what these approaches are and why they are good.

We mean that LEMs should follow hydrological modelling approaches to uncertainty generally, and indeed there are plenty. The Lisflood-FP approach is provided as an example, one which is widely used, and makes the most sense for CAESAR-Lisflood for obvious reasons. Will rephrase this paragraph to make this clearer.

523. But it is quantitative! I think you are again confusing "subjective" and "nonquantitative".

Have removed these references.

530-533. This is a bit of a bombshell that you are dropping on yourselves at the end: so you are unsure that the experimental design fairly weights the sediment transport formulas compared to the other values? There seems to be an easy answer, though: just take the binary extreme values of the other variables, and compare a subset of the runs with only 2 states considered for each parameter?

Yes, this is a bombshell, and on reflection one we have dropped on ourselves unfairly. The purpose of the test was to use the Morris Method to assess the impacts decisions on parameter values have on the behaviour of the model – in this case addressing SED as a binary choice is entirely appropriate and justified. The Morris Method is subjective, and its purpose is to guide an operator in calibrating a model by identifying which parameters impact the model the most. The choice of SED is binary in this version of CAESAR-Lisflood and has the largest impact on the model behaviours. The minimum and maximum extents of the other parameters were deliberately set wide (+/- 50 %), wider than would normally be considered in a sensitivity test (eg, recent UK Environment Agency guidance suggest varying Manning's n +/- 20% to test a model's sensitivity). In this sense we could almost argue the opposite, that the impact of the SED choice is understated.*

In response to your suggestion, instead we have reprocessed the model results assuming that the two SED formulae are the min and max choices across 5 steps. Therefore, switching from one to the other is 4 iterative step changes, and we divide all associated elementary effects by 4 and replot the normalised data. Here we see that the model shows less sensitivity to the parameter change, with others, such as Manning's n and Grain Size Set overtaking it. This could be argued that it is truer reflection of SED's role in model uncertainty relative to other parameters, but we would argue that the information because less useful to the operator as it does not reflect the true impact of the decision on the model outputs.

We had modified the discussion from Line 789 – "The main limitation of the MM is the subjectivity in selection of parameter values and ranges. Here, this has been mitigated by consistently selected ranges of +/- 50 % of a default value obtained from previous calibrations (where feasible). An issue emerges with categorical parameters, such as SED, where multiple values cannot be placed in spectrum across a range between minimum and maximum values. The MM has no formal method for dealing with such categorical parameters, so here it has been assumed that switching from one formula to another is a single iterative step change, and this would be the same even with more choices available. This reflects the purpose of the MM, which is to inform about the relative importance of choices of parameter values on the performance/behaviour of the model. However, to assess the impact of this single step-change assumption, we performed a further analysis, where it was assumed that switching formula was a change of four iterative steps. This analysis shows that the relative sensitivity of the model to the sediment transport formula choice becomes less important, with other parameters such as Manning's *n Roughness and grain size sets increasing in relative influence (see Supplementary Material S2 for full results of this analysis)."*

We have included the additional tests as supplementary material.

**Hankin, B., Arnott, S., Whiteman, M., Burgess-Gamble, L., and Rose, S., 2017. Working with Natural Processes – Using the evidence base to make the case for Natural Flood Management.* Environment Agency Report – October 2017. Project Number - SC150005

534-543: I think that the compute time and number of models should be mentioned far earlier in the paper (methods/results), and then perhaps referred to here as a reason for your decisions.

From line 817 – "The bulk of simulations used Intel i7-5960X processors and using Solid State Drives (SSD), yet the run times varied considerably depending on the parameter sets chosen. As an indication, the mean simulation run time for the first repeat in each catchment was 11 hours and 23 minutes for the Swale and 21 minutes for Tin Camp Creek."

Reviewer comments are shown in black, *and author responses are shown in red. Line numbers in responses refer to the marked-up version of the revised manuscript*

GMD Paper Review Response – Reviewer 2

Review 2 – Daniel Hobley

First of all, sincerest apologies for the extreme delay in providing this review. In this submission Skinner and colleagues present a new approach to understanding the sensitivity of landscape evolution models using the Morris method, and using model functions in place of objective functions. This approach is semi-quantitative and somewhat subjective, but nonetheless has utility in assessing sensitivity for such models where computational demands may be prohibitive for a "proper" SA. They outline the method and apply it to the CAESAR-Lisflood model being used to simulate a pair of catchments. They illustrate that the approach works in this context, and that it highlights the key importance of selection and calibration of the sediment flux law above all

other parameters. They also discuss other aspects of the model utility, using these two cases as examples. I enjoyed this manuscript; the approach seems simple, but given the dire state of past attempted SA in geomorphic modelling this is a very much worthwhile contribution to the literature. In my opinion it requires minor to moderate revision before acceptance, as detailed below. My primary concerns relate to lack of clarity in the methods. Per GMD's review criteria:

1. I believe this paper sits within the scope of GMD, though I do not feel best placed to judge this. It presents a novel approach to the sensitivity analysis of LEMs.

2. Both ideas and tools are novel.

3. The paper seems to represent a significant advance in the state of the art of sensitivity analysis within the field.

4. Assumptions are made clear, but description of methods needs further attention. As it stands, the method could not be understood in its entirety based only on the text.

5. Results support the interpretations and conclusions, assuming I have adequately followed a couple of opaque parts of the methods.

6. See 4. Significantly more methodological detail is needed.

7. Credit is given. Abstract could even put more emphasis on the "model function" aspect of this work, which seems novel and key to the approach.

8. Title describes paper

9. Abstract is concise, though needs a touch more definition of terms to make it crystal clear to the non-expert.

10. Presentation is good

11. Language is fluent and largely precise, though I have flagged up a few instances of imprecision related to the naming of model input and output information ("parameters"?).

12. Symbology is good, though the equation presented perhaps could be tweaked to enhance clarity.

13. Some clarification is necessary throughout. Structure is good.

14. References good.

15. Code is not supplied, but freely available via the net.

Thank you for the detailed and very constructive review.

However, the supplementary information is confusing, in that S2 does not appear to be referred to from the text. This needs to be thought through and resolved. Given the brevity of the paper and of S2, and the importance of the topic it discusses, it should probably be integrated into the main text. I have not attempted to formally assess the fit of this manuscript to GMD, though I believe it is appropriate. I have also not attempted to check in any way that the detailed requirements for publication in GMD (e.g. version numbers, adequate documentation) are all met, and leave this to the editor.

We have rewritten much of the Methods section to make our approaches much clearer. In particular the description of the Morris Method has been heavily edited, from line 337 –

"Our study used the MM described in Ziliani et al. (2013), i.e. the original MM of Morris (1991), as extended by Campolongo et al. (2007), and applied the "sensitivity" package in the R Statistical Environment (Pujol, 2009) to generate the parameter sets for the SA.

To set up the MM we selected a number of parameters to be assessed, specifying a minimum and maximum range for each, plus a number of iterative steps. The parameter values are equally spaced based on the range and number of steps – for example, a parameter with a range of 2 to 10 and 5 iterative steps would have available values of 2, 4, 6, 8, and 10. This is done for each parameter and, where possible, the same number of iterative steps was used for each.

The MM samples the global parameter space by performing multiple local SAs referred to as repeats. The first simulation in each repeat is made up of a randomly assigned selection of parameter values from the available values. To set up the second simulation in the repeat a single parameter is randomly selected and its value changed by a random number of iterative steps – if we use the example above, if simulation 1 used the value 4, changing this to 2 or 6 would be one iterative step change, to 8 would be two steps, and using 10 would be three steps. For simulation 3 in the repeat another randomly selected parameter is changed although previously changed parameters are no longer available to be selected. This is continued until no further parameters are available to be changed, therefore in our study each repeat contains 16 tests – 1 starting set of parameters, plus 15 parameter changes. In this study we have used 100 repeats, for a total of 1600 individual simulations – for comparison, the implementation of the MM by Ziliani et al. (2013) used 10 repeats.

The sensitivity of the model to changes in parameter values is evaluated by the changes of objective function values between sequential tests within repeats relative to the number of incremental steps the parameter value has been changed by. The change in objective function score between two sequential tests divided by the number of incremental step changes is an elementary effect (EE) of that objective function and the parameter changed (Equation 1). After all 1600 tests have been performed, the main effect (ME) for each objective function and parameter is calculated from the mean of the relevant EEs – the higher the ME the greater the model's sensitivity. Alongside the ME, the standard deviation of the EEs is also calculated as this provides an indication of the non-linearity within the model."

Most of my comments are best suited to a line-by-line approach. However, a brief overview is warranted:

* The main issue I had with the manuscript was related to lack of detail in the methods, and some sections where key concepts needed expounding on more. The inline comments detail this, but this is

essential work. It is difficult to interpret the presented results for yourself because some crucial, detailed information is missing.

We have heavily edited the Methods in line with the comments. Please see above.

* In particular, the methods are very opaque when thinking about the time component of the models, in that you basically don't talk about it. How is it determined when a model run is "done"? Is there some external constraint on total time to run for? Please provide more information.

All runs were 30 years, including a 10 year spin-up period. This is inferred in places but you are correct that this is not explicitly stated. This is now stated in Lines 272-274.

* The importance of the sediment equation choice. Skimming Andy's review, I think I largely agree with his criticisms on this front, though see below for detail in this review. I suspect a lot of this criticism is again coming from too brief a description of this part of the methods. You may be able to head off a number of my concerns simply by expanding, and taking a more pragmatic approach to why you've made these assumptions (i.e., this is how a lot of models are applied "in the wild", ignoring known geomorphic complexity, so this is how you've done it here; this is an illustrative study so almost the details of the actual geomorphology in those places don't matter; it's instructive to see if there's any influence from known imperfections in the model assumptions; etc)

We have made this clearer in the revised manuscript. From line 263 – "It is important to state that this study is an illustration of the potential for using the MM to inform an operator of how model parameter choices can impact the performance and behaviour of their model. It is not an attempt to reproduce or calibrate the CAESAR-Lisflood model to real-world observations, although the model has been applied to each catchment previously."

*I found it hard to keep some of your terminology straight, largely around which "parameters" or "metrics" you meant at points later in the text. Take more time to define things more clearly at first use, then be very careful to define those two terms as one of your other input/output classes whenever used subsequently (there's a lot of detail on this below).

We have tightened up our use of this terminology and provide a glossary of key terms from lines 280.

* A variety of stylistic/text things, though most of these a copyeditor will catch (e.g., rogue capitalisation)

Have edited throughout.

* Some concepts need to be introduced earlier in the manuscript, e.g., Morris method definition, model function.

The Morris Method is now introduced much earlier in the manuscript, from Line 138.

* I think the importance and novelty of the Model Function approach in this context is quite underplayed, and could be brought out more.

We have made the model functions more prominent throughout, including in the abstract and by introducing them earlier. From line 224 – "The paucity of observational data and the lack of measures that amalgamate the complexity of spatio-temporal landscape change into a single metric have prevented the objective function approach to be common in modelling landscape evolution. Instead, LEMs can be evaluated by observing the changes in model outputs reflective of model behaviour – these model functions can be used in lieu of objective functions to allow the sensitivity of LEMs to be assessed. Model functions would be best used as a set in combination to allow assessment across a *range of model behaviours, and would also be transferable across a range of catchments. Such an approach formalises existing methods of evaluating LEM outputs and provides a framework from which multi-criteria objective function approaches can be applied when suitable observation become available."*

In summary, I thought this was a succinct, neatly packaged study that achieved its stated objectives, and warrants publication in GMD once it has been expanded a little. I am of course happy to provide further clarification by GMD's discussion mechanism. I look forward to seeing it promoted out of discussion paper status soon. Dan Hobley

Inline comments:

Abstract: I think the briefest of introductions to objective functions, and model functions would be appropriate inside the abstract, since they are the core of the paper.

We have rewritten the abstract. From line 16 – "The evaluation and verification of Landscape Evolution Models (LEMs) has long been limited by a lack of suitable observational data and statistical measures which can fully capture the complexity of landscape changes. This lack of data limits the use of objective function based evaluation prolific in other modelling fields, and restricts the application of sensitivity analyses in the models and consequential the assessment of model uncertainties. To overcome this deficiency, a novel model function approach has been developed, with each model function representing an aspect of model behaviour, which allows for the application of sensitivity analyses. The model function approach is used to assess the relative sensitivity of the CAESAR-Lisflood LEM to a set of model parameters by applying the Morris Method sensitivity analysis for two contrasting catchments. The test revealed that for both catchments the model was most sensitive to the choice of the sediment transport formula, and that each parameter influenced model behaviours differently, with model functions relating to internal geomorphic changes responding in a different way to those relating to the sediment yields from the catchment outlet. The model functions proved useful for providing a way of evaluating the sensitivity of LEMs in the absence of data and methods for an objective function approach."

31: "dominant"

47: perhaps "from Earth surface processes, for instance, ..."

Changed, line 64

61: Surely "legitimated by theories" -> "directly physically constrained or measurable"

Changed, line 78-79

66: Probably add something to the effect of: SAs are key in scenarios where input parameters are tuned (i.e., link to the prev bit about parameterisation more firmly)

Acknowledged, line 84 – "This is useful for identifying key parameters for later calibration but this has rarely been conducted for LEMs".

74: A bit more context to set the scene here: which fields in environmental sciences have been doing this well? Can you give a couple of examples?

Acknowledged fields using these methods more clearly from line 116 – "The use of SA as a routine component of model assessment and calibration is common place in climatic, meteorological, hydrological, hydraulic and many other modelling fields."

92: capitalisation

116: “The study” – Ziliani ref has not appeared any time recently in the text. Rephrase this bit, and set the context for why you’re about to discuss this specific study.

Changed at line 155 following the earlier introduction of MM.

118: Perhaps “see below” for the MM Section 1.1 – I found this to be very clear.

130-131: Grammar issue to do with word “both”

127-141: Another major issue is that even with data in hand, it is challenging to derive meaningful metrics from them. E.g., a pixel-by-pixel map of topography is not a good optimisation target, since small error in the exact position of the channel gets magnified when in fact the model may be doing a good job overall. So integrative metrics must be found, and that in turn is subjective. There’s also the related issue of stochastic processes inside landscape response being hard to capture in a spatially resolved way; see, e.g., work by Mary Hill (I see this gets discussed obliquely around ln 150, but be a bit clearer about it – i.e., give us the logic by which Hancock & Wilgoose (ln 154) started using statistical measures in the first place).

This is a very good point. We think as model operators we seek a single catch-all measurement from which we can derive an objective function. LEMs are likely to be more spatially varied (or rather spatial variations have a greater impact) than hydrological models for example, so this one single measure probably does not exist. Instead, we ought to pursue multi-criteria approaches as objective functions.

We have made numerous small edits to this section to reflect these points. Notably from Line 209 – “Finally, the spatially and temporally heterogeneous response of erosion and deposition patterns in LEMs also makes “pixel-to-pixel” comparisons difficult. For example, in a valley reach, gross patterns of erosion and deposition may be identical but with the channel on the other side of the valley – yielding a poor pixel-to-pixel comparison.”

157: “the LEM SIBERIA”

This reference has been removed

160-161: “these measurements” not clear which measurements. Rephrase.

This has been removed in the rewriting of this section.

Ln 164: For me, “/reliability” is redundant here. If I was you, steer clear of the implication that an accurate model is necessarily good for future prediction, as that’s a slightly different thing.

As above.

172: “MM” You need to introduce the Morris Method in the intro somewhere to say why it’s important, who has tried using it before, etc etc. Calling forward to the methods for most of it is fine, but the intro needs a little more.

Now introduced much earlier, at line 138.

172: Dash is probably a new para.

175: As for the MM, the intro also needs to define a “model function” before you get here. (Also, consistency of capitalisation!)

Model functions also introduced much earlier. From line 224 – "The paucity of observational data and the lack of measures that amalgamate the complexity of spatio-temporal landscape change into a single metric have prevented the objective function approach to be common in modelling landscape evolution. Instead, LEMs can be evaluated by observing the changes in model outputs reflective of model behaviour – these model functions can be used in lieu of objective functions to allow the sensitivity of LEMs to be assessed. Model functions would be best used as a set in combination to allow assessment across a range of model behaviours, and would also be transferable across a range of catchments. Such an approach formalises existing methods of evaluating LEM outputs and provides a framework from which multi-criteria objective function approaches can be applied when suitable observation become available."

Ln 184: As above, please deal with the MM definition before we get here.

Line 138.

195-6: This needs a supporting reference.

Have removed this statement.

2.2 – a big chunk of this stuff should be sat in the intro, as it introduces the method and its background.

This has been moved to the introduction – line 138.

244: "stochastically"->"randomly"?

246-7: "a number": how many? Is this constant? Do we step around the values in a random-walk fashion, or is this more like a ratchet to move through all possibilities? i.e., more detail here.

The description of the MM implementation has been rewritten.

250: the Main Effect can't be both the mean and SD of the EEs, as the text here implies as written. Presumably it's the mean, but make it clear. The role and handling of the SD in general in this method is unclear, partly due to folding it into the ME here then talking later about parameter normalisation against the ME. This makes it difficult to parse whether the SD is normalised or not, and if not, why not. (cf lns 362, 377->, fig 3) Expand and take more time over the details here.

Included in revised section, from line 365 – "After all 1600 tests have been performed, the main effect (ME) for each objective function and parameter is calculated from the mean of the relevant EEs – the higher the ME the greater the model's sensitivity. Alongside the ME, the standard deviation of the EEs is also calculated as this provides an indication of the non-linearity within the model."

ln 255: something not right with this equation – it doesn't read sensibly. Neither r nor j appears in the RHS of the equation. Feels like there should be subscript outside the bar symbology indicating how we move through j space. Check the equation, consider again if this is the clearest way you can show it, and be considerably more generous with the explanation. The fact that an EE is parameterised as "d" seems somewhat perverse... This would be a lot clearer if you gave a concrete example here, e.g., say what the data structure would look like for a given model output measure *and* Model Function. It's very opaque how model output measures and model functions differ. Make this explicit both here and especially later in 2.5.

This is not a new equation nor is it our equation – we present it here as it has appeared in previous literature and would not wish to change terms, such as "d". However, we have clarified our own description as it was unclear (lines 406-409).

It calculates the EE, as the change in the model function value (i), when a single parameter value is changed (j is the parameter changed). Where "model output measure" is used, it should read "model function". In other words, it calculates the change in function score between one test and the next one within a repeat.

271: repeated how? Presumably there are other inputs to the model besides precip (e.g., land use, topo, etc etc) so either talk about them all here, or none of them. What's the source of the DEM in each case?

Included more details in the revised manuscript.

2.3.2: as at 271; more of this info is presented here, but it still needs more.

More details included..

2.3.1 & 2.3.2: What e.g. Shrahler order are these streams? This needs saying to present a contrast with your different order measure, lns 285-287.

See below (285-7)

Fig 1 caption: Add "Note difference in scale"

285-7: What is the justification for this division, which, AFAIK, is novel? Explain why an existing stream order method is not chosen.

This is a novel method. Existing methods, eg Strahler stream numbers, can result in different order numbers due to the connectivity of the catchment at a particular moment, and can also be influenced by the resolution of the DEM – eg, smaller stream orders would not be picked out on the relatively coarse 50m grid used for the Swale. The new method was used to compare to two very different catchments and DEMs in a consistent way, and no further claims are made to its physical basis or usefulness. Included an acknowledgement of this line 447 – "This method is novel and was developed to provide a consistent method of sub-dividing both catchments independent of factors such as connectivity and DEM resolution."

303-304: "excluding... dune and soil development". Return to 2.1 and explain in more detail which processes are turned on and off in CAESAR-Lisflood. (Note this is not the same thing as which are being tested as part of the SA.)

From line 274 – "CAESAR-Lisflood model is used in catchment mode, the simulations have no representation of suspended sediments and bed rock, and the dune and soil evolution modules are not used."

305: Please give information on the typical run time of each model run on whatever rig was used to give the reader some context on what this information means.

From line 817 – "The bulk of simulations used Intel i7-5960X processors and using Solid State Drives (SSD), yet the run times varied considerably depending on the parameter sets chosen. As an indication, the mean simulation run time for the first repeat in each catchment was 11 hours and 23 minutes for the Swale and 21 minutes for Tin Camp Creek."

Ln 311: I believe the fragment about it being a qualitative method is unnecessary here- delete it. I'm not sure it is technically true, and not needed anyway.

Have edited this, from line 479 – "MM is subjective in that the relative sensitivities shown depend on the minimum and maximum range values set by the user."

312-318: "it's difficult to define what this means" I agree...! I get the gist of this, but you need to tighten it up and expand significantly. Delete the sentence about "broadly equal" – which isn't true, since I assume the values themselves are not equal, and say what you actually mean ("to have an approximately equal influence on the output, as defined by...?" Or something??) If you get it right, you can and should remove the "It is difficult to define what this means".

Have rephrased this section, from lines 479-486.

Ln 316 – On what basis is it sometimes not appropriate to do this?

Lines 484-486 – "(for example, the Manning's n Roughness for Tin Camp Creek where +/- 50 % would have resulted in obviously physically unrealistic values – see Table 1 for values used)."

Ln 317/8 – give us the specifics of the steps you used for Manning's n, to make this more concrete.

All the stepped values are presented in Table 1. Included reference in line 486.

320-22: This could do with a lot of expansion, as it becomes key to the results and discussion. On what basis were these two laws selected from the large field of possibilities? The second sentence needs a lot more detail. This is the first time we're hearing that the MM can be used to tackle epistemic uncertainty in model set-up, rather than common-or-garden parameter uncertainty. Tell us a lot more about how this works and why you're doing it, then explain with more detail what exactly "as binary two-step" means. I presume you mean the choice of law becomes a parameter in the MM, but in that case, aren't you also switching in and out another big subset of the parameters? So how is your data now comparable? Doesn't the parameter set have to stay static for the MM to work? I may have misunderstood, but if I have, it's a sign you need lots more detail on how the MM works.

Simple answer to this is that these are the two sediment transport formulae available in the downloadable model, and therefore those available to a typical operator without modifying the model code.

Line 489 – "These were not selected as representing the best fit for the catchments simulated but because they are the formulae available in the unmodified version of CAESAR-Lisflood."

We apply the choice of one or the other as a parameter, so in the Morris Method it will switch between them once in each repeat. When it changes no other parameters in the model are changed at all. We will make this clearer when we rework the methods section.

Line 492 – "(no other parameter values were varied when this occurs, as per the description of MM in Section 2.2)."

327/8: Formatting of ref. Otherwise, I like this GS description.

Ln340-344: This could be clearer. Surely some statistical methods have promise; after all, you're about to devise one. You need something in here to refine the scope of the methods you're talking about, e.g., do you just mean to exclude previously attempted methods? Is it the fact that single discrimination criteria don't work, and that's the problem? Etc

Yes, this was not clear. We have rephrased this section between lines 523-534 – "The common method of assessing a model's sensitivity to parameters values via SA, and the method employed by the MM, is to observe the variations to objective function measures. However, the difficulties in applying an *objective function approach to LEMs were highlighted in Section 1.2, and in order to apply an SA a novel approach is required. The method we have developed eschews the objective function approach and instead assesses the model against a series of model functions designed to reflect some of the core behaviours displayed in the model – these can be seen in Table 2. This represents a philosophical difference to traditional applications of SA – here we are not testing the model against its skill in simulating the physical environment, but rather how the model responds behaviourally to changes in the user-defined parameters detailed in Section 2.4. The 15 model functions (Table 2) are simple, scalable and transferable between different catchment types, and can be applied to simulations of different timeframes. The model functions are based on outputs which are not unique to CAESAR-Lisflood, so can be applied to other LEM and geomorphic models."*

Ln 344: say why these reviewed methods failed, since referring out the 2001 work doesn't provide the context.

This reference is no longer included in the revised section.

Ln 349: You need to be clearer in defining that Model Functions are a new thing that you're creating in this work, and that this is a major contribution of this work. The introduction should reflect this (and thoroughly introduce the idea), as should the abstract. If of course it isn't your idea and it's coming from somewhere else, please make that clearer and add the relevant references more clearly. Also, as noted earlier, please explicitly discuss more clearly how model output measures differ from/work with/are part of Model Functions (See also 360-2).

Have included more prominent references to the model function throughout, including in the revised Abstract and the Introduction.

Ln 353: dash is a sentence break

This section has been revised.

360-362. Rephrase; this doesn't fully make sense. Do you really mean "versus"? I think part of this needs to read "was normalised to the highest ME for any parameter within the Model Function". Again, it's really hard to follow and keep straight the respective roles and interrelationships of: the parameters back in the equation (i.e., i, j), vs what you mean by "parameters", vs Model Functions, vs model output measures, vs MEs, etc etc. I'm fairly some of these are equivalent to each other, but it's very hard to keep straight, even when flicking back and forward back to the methods. Make your terminology bulletproof, crisply defined, and repeat yourself as needed for maximum clarity – as this is very hard to follow. As above, illustration by example would make this a lot clearer.

We hope the inclusion of the glossary of terms and the tightening of our use of terms have improved the clarity of this section.

362/364: "aggregated". I'm conceptually uncomfortable with taking a mean (right? That's what you mean?) of numbers which have already been turned into the equivalent of percentages and scaled to each other. This will create some odd statistical dynamics, I think. I guess this is fine given the method is already pretty qualitative, but it makes me uneasy. I'd invite the authors to reflect on this, and consider putting something in to reassure the sceptical reader. More concretely, how are you aggregating standard deviation measurements? Are these scaled alongside the MEs? If so, that will very quickly get confusing. Explicitly tell us what you are doing to handle these, and let the reader sort it out for themselves (cf, ln 250).

Yes, the aggregated values are the means of the normalised ME and the means of the normalised standard deviations of the EEs. This was used to illustrate the large amount of information we had and admittedly is far from perfect – it does provide a useful, concise summary of which parameters influence the model the most across a range of model functions. We will edit the text to make it clear that this is illustrative and that operators using this method should rely on the results for each individual model function (it is unusual for the Morris Method to be used to assess so many different functions).

Have included a caveat in line 560 – "To summarise the large amount of information produced, the ME of each parameter and model function combination was normalised based on the proportion of the ME for highest ranking parameter for that model function"

Have included description of aggregating the SD, line 565 – "The same was also done, separately, for the standard deviations of each parameter and model function."

365-371: I think this means that, implicitly, these methods won't work to compare true transient behaviours? It's impossible to understand without telling us how the method treats time in general. Which leads to... General comment: In the methods, the role of time in this approach is very unclear. Time elapsed is never referred to until you're talking about trimming off spinup. Is EE calculated continuously through time, and you stop at the best possible value? Are you comparing time series? Or just best fit at a time slice? This information must be present. In general, this is a symptom of your description of the key numerical methods being very brief. Be much more generous, with examples etc. as you work through from EEs to ME to Model Function aggregates. Take as much space as you need and let this "breathe" as much as you can.

The tests were 30 years in length, but the first 10 years were excluded from calculating the model functions – this is stated in line 272.

Line 546 – "Model function values were calculated at the end of each simulation."

Figure 3: I'm concerned (possibly unnecessarily?) about cross-correlation of the means and SDs here (cf, ln 362). If you normalise the SD's by the maximum ME, then you're building in dependence of one on the other. If you don't, how did you aggregate those SDs? Are these means of means of SDs for each EE cluster, or are you taking statistical measures of ME distributions themselves? You need more methods clarity to make this easy to understand and intuitive to interpret.

The normalised values were produced separately – one for the main effects and one for the standard deviations, stated in line 565.

Fig5: please add the abbreviation to each graph caption. Each sub-graph also needs a label a, b, c etc. Although the caption specifies it, a heading on each column of graphs giving the location would make the figure easier to read. Adding "(biased smaller)" and "(biased larger)" annotations above 1-2 and 4-5 for the GSS graph would prevent the reader having to flick back a number of pages seeking the description of what these are.

Have reproduced Figure 5 accordingly.

3.3: On what basis were SED, CVS and GSS selected for this section? Justify.

Illustrative purposes only – these were the most interesting ones to us so we chose them to show.

425: This is too terse to make it clear. Expand & rephrase to be more precise about what you mean. It's not clear where you are talking about the all the figures in the figure at once, when just one site vs the other, or when you mean variation between parameters in patterns shared across both sites.

Have rewritten this section, from line 652 – "Figure 5 illustrates how changes in parameter values might influence the spatial patterns of landscape change using SED, VEG and GSS as examples. For SED (Fig 5.A and 5.B), the most obvious difference is the scale of changes seen using each formula with Einstein generally showing greater change. For Tin Camp Creek (Fig 5.B) the spatial changes are similar, but for the larger Swale (Fig 5.A) there are differences in relative rates in 2nd and 4th order areas. In the Swale, VEG (Fig 5.C) appears to have little impact on the patterns and scale of changes, yet in Tin Camp Creek (Fig 5.D) there is reduction in the rates of erosion across the catchment with higher values, except in the 5th order areas which remain at a similar level. Finally, both catchments show a reduction in rates of erosion with a greater proportion of larger grain sizes, yet this is more pronounced 4th order areas in Tin Camp Creek (Fig 5.F)."

General note: The subsections in section 4 are formatted differently to everywhere else.

Have corrected this.

431: "The implications... have been discussed". I don't feel like I've seen a discussion of the implications; the methods were very factual (rightly). You need to re-summarise whatever implications you are thinking of here. If you mean results, say so – though this would seem somewhat redundant.

Have removed this.

435/7: "metrics". Your terminology has been imprecise throughout – we can now add "metrics" to "parameters", "Model Functions", "model output measures" and "MEs" as a list of terms it's really easy to lose track of and mix up. I would advise you to remove all mention of metrics and parameters from the manuscript, or at least define which thing you mean by it each time. So here, I think you mean model output measures, right? Get everything internally consistent.

Have changed this to Model Functions (line 677)

450: "basin's". Nice points here.

463-5: "nonstationary". If I've followed (cc past comments on approach taken through time), I don't think the approach described in this paper as currently deployed would be able to analyse this kind of nonstationarity. You should say that explicitly here, though the point is well taken.

No, this method doesn't, but is a step towards using ensemble parameter sets which can in some part address this. Have removed this section entirely.

4.3: Again, points well taken, but this is more a literature discussion. This should be grounded more in your specific results. You've shown that this is the most important choice you can make in CAESAR-LF, so what does that mean? Should users be calibrating as for e.g. Siberia? Is there reason to expect this overwhelming importance to be carried over to other sites? Does this mean calibrating SED well is by far the most important thing, so other effects are secondary? This section is negative/bear-ish, and doesn't outline any actual approaches, advice, or other opportunities to use the SA productively to get around this problem. How bad actually is the lack of constraint on SED? Even a statement along the lines of "this analysis suggests that detailed justification and calibration of model choices around sediment transport will lead to the most effective gains in model skill" would help. This seems like a key result here.

This is difficult – as deterministic sediment transport laws applied spatially over a LEM grid yield can result in chaotic responses (erosion/deposition stores, the ability to supply limit sediment etc.). A LEM CAN be calibrated in a fashion – but given the chaotic response of the sediment system this can only be general. And then will also be subject to the difficulty of calibrating only to the events (and catchment history) of the calibration sequence. The response of the sediment system to changes in rainfall is exponential (see Coulthard et al., 2012) so even small increases in RF inputs outside of the calibration set may give disproportionately large outputs (or not as its chaotic!). A further issue for calibration is the lack of field data with which to calibrate LEMs.

We have divided this section into two in the revised manuscript, one dealing with the role of sediment transport formula in LEMs, and another considering the implications for the calibration of such models (at line 737).

From line 750 – "Furthermore, this analysis suggests that detailed justification and calibration of model choices around sediment transport will lead to the most effective gains in model skill."

And leading on from this -> Ln 528: "binary". I also have some issues here regarding the implementation of the SED equations – you say the choice was binary, but how then did you calibrate the internal parameters in each equation (cf, comments in methods section above)? This might be resolved by more clarity in the methods, but also, it seems like maybe you could have got inside this element of the model a bit more and explored those subparameters too, per Andy's comments. Are you simply applying the different equation with the same parameters in each? Or do you allow some kind of pre-optimisation to get the params to where they need to be in each field site? Additional material in the methods clarifying this is essential, and I'd also strongly advise you to refer explicitly to how this choice of sed law parameterisation is affecting the importance of the SED param, both here and in 4.3 above. I see material in S2 discusses this, yet isn't referenced from the text. I'd advocate pulling that material into the main text as part of addressing this comment.

We have explained this above where you previously raised this – when the Morris Method switches from one formula to the other, all other parameter values remain constant. There is no optimisation as we made no attempt to optimise the model (indeed, there are no observation data to even try!).

Conclusion: please switch out MNR, GSS and VEG, IOD, etc etc for text equivalents to make this conclusion more stand-alone. I also feel like the influence of SED is a key result, and should also be mentioned in here. Otherwise, I like this as a summary.

Have made these edits.

Global Sensitivity Analysis of Parameter Uncertainty in Landscape Evolution Models

Christopher J. Skinner[1], Tom J. Coulthard[1], Wolfgang Schwanghart[2], Marco J. Van De Wiel[3], and Greg Hancock[4]

[1]School of Environmental Sciences, University of Hull, Hull, UK

[2]Institute of Earth and Environmental Science, Potsdam University, Potsdam-Golm, Germany

[3]Centre for Agroecology, Water and Resilience, Coventry University, Coventry, UK

~~[1]University of Hull, Hull, UK~~

~~[2]Postdam University, Postdam, Germany~~

~~[3]Coventry University, Coventry, UK~~

[4]University of Newcastle, Callaghan, Australia

Corresponding Author: C. J. Skinner (c.skinner@hull.ac.uk)

Abstract

The evaluation and verification of Landscape Evolution Models (LEMs) has long been limited by a lack of suitable observational data and statistical measures which can fully capture the complexity of landscape changes. This lack of data limits the use of objective function based evaluation prolific in other modelling fields, and restricts the application of sensitivity analyses in the models and consequential the assessment of model uncertainties. To overcome this deficiency, a novel model function approach has been developed, with each model function representing an aspect of model behaviour, which allows for the application of sensitivity analyses. The model function approach is used to assess the relative sensitivity of the CAESAR-Lisflood LEM to a set of model parameters by applying the Morris Method sensitivity analysis for two contrasting catchments. The test revealed that for both catchments the model was most sensitive to the choice of the sediment transport formula, and that each parameter influenced model behaviours differently, with model functions relating to

internal geomorphic changes responding in a different way to those relating to the sediment yields from the catchment outlet. The model functions proved useful for providing a way of evaluating the sensitivity of LEMs in the absence of data and methods for an objective function approach.

~~Landscape Evolution Models have a long history of use as exploratory models, providing greater understanding of the role large scale processes have on the long term development of the Earth's surface. As computational power has advanced so has the development and sophistication of these models. This has seen them applied at increasingly smaller scale and shorter term simulations at greater detail. However, this has not gone hand in hand with more rigorous verifications that are commonplace in the applications of other types of environmental models- for example Sensitivity Analyses.~~

~~This can be attributed to a paucity of data and methods available in order to calibrate, validate and verify the models, and also to the extra complexity Landscape Evolution Models represent - without these it is not possible to produce a reliable Objective Function against which model performance can be judged. To overcome this deficiency, we present a set of Model Functions - each representing an aspect of model behaviour - and use these to assess the relative sensitivity of a Landscape Evolution Model (CAESAR-Lisflood) to a large set of parameters via a global Sensitivity Analysis using the Morris Method. This novel combination of behavioural Model Functions and the Morris Method provides insight into which parameters are the greatest source of uncertainty in the model, and which have the greatest influence over different model behaviours. The method was repeated over two different catchments, showing that across both catchments and across most model behaviours the choice of Sediment Transport formula was the dominate source of uncertainty in the CAESAR-Lisflood model, although there were some differences between the two catchments. Crucially, different parameters influenced the model behaviours in different ways, with Model Functions related to internal geomorphic changes responding in different ways to those related to sediment yields from the catchment outlet.~~

~~This method of behavioural sensitivity analysis provides a useful method of assessing the performance of Landscape Evolution Models in the absence of data and methods for an Objective Function approach.~~

1. Introduction

Landscape Evolution Models (LEMs) investigate how the Earth's surface evolves over timescales ranging from hundreds to millions of years (Coulthard and Van De Wiel, 2012; Martin and Church, 2004; Pazzaglia, 2003; Tucker and Hancock, 2010; Van De Wiel et al., 2011). They represent the earth's surface with a regular or irregular mesh and simulate how the surface evolves over time as a function of tectonic processes, and erosion and deposition from ~~fluvial, glacial, aeolian and hillslope~~Earth surface processes. LEMs have proved to be very useful scientific tools to understand how Earth surface processes interact to shape the landscape.

More recently, LEMs have improved considerably in their ability to simulate the physical environment, and this has developed in parallel with improvements in computational efficiency and power. This allows LEMs to go beyond highly simplified models of landform development ~~but~~ and to also incorporate increasingly complex processes such as pedogenesis (Vanwalleghem et al., 2013; Welivitiya et al., 2016) and periglacial processes (Andersen et al., 2015; Egholm et al., 2015). Other processes are now being handled in more detail such as hydrodynamic flow models and aeolian processes (Adams et al., 2017; Coulthard et al., 2013; Liu and Coulthard, 2017). These developments led to Coulthard et al. (2013) describing them as 'second generation' LEMs that extend previously explanatory and explorative models to be used for prediction of future changes in landscapes, such as for the mining industry (e.g. Hancock et al., 2017; Saynor et al., 2012).

However, more detailed physical representations of the processes that shape the Earth's surface involve a larger number of parameters that are typically estimated from proxy data or theoretical considerations, or are completely unknown ~~not legitimated by theories but must be determined from empirical data or are incompletely known~~ (Oreskes et al., 1994; Petersen, 2012). If LEMs are to be operationally used for prediction or as decision-making tools in the future, their outputs must be evaluated against the uncertainty in input parameters – a task that is increasingly difficult for a large number of parameters. ~~Sensitivity~~ Through sensitivity analysis (SA) investigates how variations in the output of a numerical model can be attributed to its input factors (Pianosi et al., 2016).~~,~~ This is useful for identifying key parameters for later calibration but this has rarely been conducted for LEMs. The aim of this study is thus to conduct a SA of the widely used and highly parameterized LEM CAESAR-Lisflood (Coulthard et al., 2013) - in particular, we wish to be able to detect the parameters that have the greatest influence on the model's simulation output. As model sensitivity may be influenced by different landscapes, we run the SA in two individual and distinct catchments.

1.1 Sensitivity Analysis and Landscape Evolution Models

The application of SA in environmental modelling has a history spanning four decades (Norton, 2008) and forms an important component of using models for decision-making, including model development, calibration and uncertainty analysis (Yang, 2011). SA addresses five key questions (Cariboni et al., 2007; Neumann, 2012; Song et al., 2012, 2015):

1. Which parameters have the greatest influence on the model?
2. If additional data could be used to reduce the uncertainty in a parameter, which would most reduce the model output variance?
3. Are there parameters with such low influence that their values could be fixed without impact on the model outputs?

4. If parameter values emerge as incorrect, how will they influence model outputs?
5. Which parameters influence model outputs in different regions (parameter space)?

Clearly, based on the above, an appraisal of model sensitivity is important to fully understand and apply model results. In a review of applications of SA in environmental models, Yang (2011) identified two common approaches to SA – local and global. Local SA are limited, considering only the impacts of factors on model outputs locally, i.e. within a restricted region of the model's parameter space, whilst global SA typically utilise Monte-Carlo methods to assess the sensitivity of impacts across the whole parameter space (Yang, 2011). For complex models with non-linear behaviours, the use of Local SA can be highly biased as they neglect the non-linear interactions between parameters (Oakley and O'Hagan, 2004; Pappenberger et al., 2006; Yang, 2011). Global SA are more computationally expensive, but as the methods are more reliable, they are attractive to modellers (Yang, 2011).

The use of SA as a routine component of model assessment and calibration is common place in climatic, meteorological, hydrological, hydraulic and many other modelling fields. However, for LEMs there are surprisingly few examples of SA being carried out. This can be explained by three inter-related issues: (i) LEMs typically have a large number of model parameters; (ii) long model run times can make multiple simulations for SA impractical; and (iii) model behaviour can b~~h~~e highly non-linear (e.g. Coulthard and Van De Wiel, 2007; Larsen et al., 2014; Van De Wiel and Coulthard, 2010), leading to potentially complex SA interpretations. Large numbers of model parameters and long run times, in particular, make Monte-Carlo methods extremely time consuming – and therefore often unviable.

There are several ~~past~~ studies ~~investigating~~ on how LEMs respond to variable forcing, process changes and model parameters, including changes in climate variability and precipitation resolution (Armitage et al., 2017; Coulthard and Skinner, 2016a; Ijjasz-Vasquez et al., 1992; Tucker and Bras, 2000), channel widths (Attal et al., 2008), vegetation (Collins, 2004; Istanbulluoglu and Bras, 2005), and variations in initial conditions (Hancock, 2006; Hancock et al., 2016; Ijjasz-Vasquez et al., 1992; Willgoose et al., 2003). Campforts et al. (2016) investigated how different numerical solvers affect LEM simulation. Yet few studies explicitly perform SA and most of the applications described above are exploring LEM sensitivity to processes, or changes in environmental conditions, and are more correctly referred to as exploratory tests (Larsen et al., 2014). On the other hand, investigations to ascertain the model's response to potential uncertainties (e.g from model parameterisation) can be deemed as true SA (eg, Armitage et al., 2017; Coulthard and Skinner, 2016a; Hancock et al., 2016).

Hydrological models faced similar issues to LEMs in the past, i.e. model complexity and long processing times when applying SA. To overcome them, hydrologists have used the Morris Method (MM; Morris, 1991). The MM can be regarded as a global SA, although it actually performs multiple local SAs sampled from across the full parameter space – this produces a series of local evaluations, the mean of which is an approximation of the global variance (van Griensven et al., 2006; Norton, 2009; Saltelli et al., 2000). The main strength of the MM is its computational efficiency. Herman et al. (2013) showed that the MM could estimate similar variance in model outputs to the Sobol' Variance-based global SA method (Sobol', 2001), yet required 300 times less evaluations, and significant less data storage for an application to a distributed catchment hydrological model. The robustness of this approach has been further shown by numerous workers (e.g. Brockmann and Morgenroth, 2007; Pappenberger et al., 2008; Yang, 2011). However, the MM cannot provide a full quantitative assessment of parameter sensitivity and is dependent upon the user-defined bounds to the parameter space. It can successfully rank parameters between the least and most influential to model outputs, but cannot determine parameters' exact relative influence (Brockmann and Morgenroth, 2007). These advantages and limitations entail that MM has primarily been used during the pre-screening stage of models, isolating the most influential parameters for further SA with quantitative, yet more computationally expensive, methods (e.g. Ratto et al., 2007; Song et al., 2015; Yang, 2011; Ziliani et al., 2013).

Field Code Changed

Field Code Changed

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

(Ziliani et al., 2013) performed a two-stage SA for the CAESAR LEM, utilising the MM (as adapted by (Campolongo et al., 2007)~~Campolongo et al., 2007)~~. Whilst this study demonstrated the feasibility of applying the MM as a global SA to a reach-scale LEM, it was applied as a pre-screening stage to identify the most relevant parameters for model calibration. In contrast, our study focuses on SA as a tool to investigate parameter influence on model behaviour.

~~The study by Ziliani et al. (2013) is another example of a LEM SA, seeking to spatially calibrate a reach-scale application of the CAESAR LEM to field observations. They performed a two-stage SA, utilising the Morris Method (MM) (as adapted by Campolongo et al., 2007) as a pre-screening before a more complex local SA was applied. The study investigated the model's sensitivity to 12 user-defined parameters, using MM to exclude those showing the least influence on performance measures from the subsequent SA and calibration. Whilst Ziliani et al. (2013) demonstrated the feasibility of applying MM as a global SA to a reach-scale LEM, it was applied as a pre-screening stage to identify parameters to focus model calibration on, and not to observe model behaviour.~~

1.2 Metrics for Landscape Evolution Model Assessment

Evaluating LEMs is challenged by the paucity of comprehensive field data against which they can be assessed and the lack of measures for calibration and validation (Hancock et al., 2016; Hancock and Willgoose, 2001; Tucker and Hancock, 2010). Moreover, some second-generation LEMs (e.g. CAESAR-Lisflood) simulate short (annual to decadal) and long-term (millennial time scales and longer) landscape changes, necessitating data and methods to assess them across variable time scales. Thus, while SA of environmental models often rely on objective functions (e.g. the Nash-Sutcliffe score between observed and simulated values; (Nash and Sutcliffe, 1970)~~)~~, this approach is generally not practical for LEMs. With few exceptions (e.g. (Ziliani et al., 2013), results from LEMs are therefore frequently assessed qualitatively, relying on visual interpretation of the simulated landforms or cross-section profiles(e.g. Coulthard and Skinner, 2016b; Hancock et al., 2010, 2015; Hancock and Coulthard, 2012).

~~An issue with the testing of LEMs is finding the field data and statistical tools that can actively discriminate between what is a good model and a bad model, and for parameterisation (Hancock and Willgoose, 2001; Hancock et al., 2016; Tucker and Hancock, 2010). As the models are designed to assess both short (annual to decadal) to long-term (geological time scale), the data and assessment methods require both a multi-dimensional approach. The application of SA to environmental models often assesses the impacts of factors based on variations in values of an objective function, which is often an error score between observed and simulated values – for example, a common approach in hydrology would to use the Nash-Sutcliffe score (Nash and Sutcliffe, 1970) as an objective function, and catchment discharges as a value. The objective function approach was used by Ziliani et al. (2013), matching the outputs of a reach simulation in CAESAR to observed patterns of wet/dry pixels, erosion/deposition, and vegetation. However, the objective function approach is generally not practical for LEMs due to a paucity of observed data to use as a value, so often the results from LEMs are assessed qualitatively, relying on visual interpretation of the final simulated landforms or cross-section profiles (eg. Hancock et al., 2010; 2015; Hancock and Coulthard, 2012; Coulthard and Skinner, 2016a).~~

~~The use of ~~C~~c~~atchment outlet statistics, such as sediment yield time series, allow for comparison between simulations to indicate a catchment's response to perturbations (e.g. Coulthard et al., 2012; Coulthard and Skinner, 2016b; Hancock and Coulthard, 2012). However, sediment yield time series rarely provide a sufficiently complete picture of a catchment's geomorphic response.~~although this provides some information about the catchment response as it gives an incomplete picture.~~ For example, Coulthard and Skinner (2016b) showed that simulations calibrated to provide equivalent sediment yields~~, to compensate for loss of spatial and temporal resolution in rainfall inputs,~~ produced different ~~landscape shapes~~landforms. For planning purposes these internal catchment changes are likely to be more useful than catchment sediment yields. Moreover, changing topography potentially instigates a feedback process that leads to complex, often non-linear catchment behaviour (Coulthard and Van De Wiel, 2007, 2013; Hancock et al., 2016; Jerolmack and Paola, 2010; Van De Wiel and Coulthard, 2010). Finally, the spatially and temporally heterogeneous response of erosion and deposition patterns in LEMs also makes "pixel-to-pixel" comparisons difficult. For example, in a valley reach, gross patterns of erosion and deposition may be identical but with the channel on the other side of the valley – yielding a poor pixel-to-pixel comparison.~~Statistics based on measurements from the catchment outlet cannot account for factors such as geomorphic equifinality, self organised criticality, and autogenics, which act as a non-linear filter on the response (Coulthard and Van De Wiel, 2007, 2013; Hancock et al., 2016; Jerolmack and Paola, 2010; Van De Wiel and Coulthard, 2010).~~

Few studies have tested metrics to compare topographic data or physical experiments to simulated elevation changes by LEM (Hancock et al., 2010, 2011; Hancock and Willgoose, 2001; Ibbitt et al., 1999). However, although the metrics often suggested a good agreement, visual analysis of the final DEMs indicated clear differences between the physical models and the simulations (Hancock and Willgoose, 2001). There is, therefore, a clear need for better statistical methods for critically evaluating and comparing landscapes that can also be used for evaluating the accuracy (or otherwise) of LEMs.

Formatted: Font: 11 pt

The paucity of observational data and the lack of measures that amalgamate the complexity of spatio-temporal landscape change into a single metric have prevented the objective function approach to be common in modelling landscape evolution. Instead, LEMs can be evaluated by observing the changes in model outputs reflective of model behaviour – these model functions can be used in lieu of objective functions to allow the sensitivity of LEMs to be assessed. Model functions would be best used as a set in combination to allow assessment across a range of model behaviours, and would also be transferable across a range of catchments. Such an approach formalises existing methods of evaluating LEM outputs and provides a framework from which multi-criteria objective function approaches can be applied when suitable observation become available.

~~Hancock and Willgoose (2001) reviewed statistical attempts to define catchment geomorphology, including width function, cumulative area distribution, area-slope relationship, and hypsometric curve, and used these as an objective function between physical experiments and numerical experiments using SIBERIA. However, although statistically similar, there were visually clear differences between the physical models and the simulations. Other methods employed include changes to mean elevations (Hancock et al., 2010, 2011), and Optimal Channel Network (Ibbitt et al., 1999). However, although visual difference may be observed between simulations, variations within these measurements have proved to be small for timescales of 1000 years and less (Hancock et al., 2010, 2011), so are limited in their scalability. There is, therefore, a clear need for more objective statistical methods for critically evaluating and comparing landscapes that can also be used for evaluating the accuracy/reliability (or otherwise) of LEMs. Field data at the catchment scale that includes erosion and deposition data, vegetation type and change as well as sediment transfer at critical points along the stream network is required. Such all-encompassing catchment scale data is currently not available.~~

1.3 A Global SA for a catchment LEM

This study demonstrates the first application of a g~~G~~lobal SA illustrate parameter influence on model behaviour in~~to~~ a catchment LEM (CAESAR-Lisflood), using the MM to assess the model's sensitivity to user-defined parameters. ~~– in total 15 parameters are selected based on~~We selected 15 model parameters chosen either because of their known importance to the model or because the model's response to the parameter is presently poorly understood. Although not all the 15 model parameters ~~chosen~~ are universal between LEMs, many LEMs have equivalents.~~ A~~ Moreover, we developed a set of 15 model functions ~~has been developed which reflects~~ that reflect core behavioural responses of the model.~~,~~ ~~and these~~These will indicate whether the same parameters influence all behaviours, or whether the different behaviours respond to different parameters. The choice of 15 model parameters and 15 model functions is coincidental. ~~The method is applied to~~We conducted the SA in two ~~contrasting~~ catchments ~~(scale, environment and climate)~~with contrasting environmental settings to assess how transferable an individual SA is to different conditions.

It is important to state that this study is an illustration of the potential for using the MM to inform an operator of how model parameter choices can impact the performance and behaviour of their model. It is not an attempt to reproduce or calibrate the CAESAR-Lisflood model to real-world observations, although the model has been applied to each catchment previously.

2. Methods

We apply the MM to perform a global SA on the CAESAR-Lisflood model for two contrasting catchments (more detail in Section 2.3): the Upper Swale, UK (181 km^2, temperate, perennial), and Tin Camp Creek, Australia (0.5 km^2, tropical, ephemeral). Each individual simulation runs for a 30 year period, where the first 10 years are used as a spin-up to reduce the impacts of transient model behaviour and therefore output analysis starts after year 10 of the simulation. CAESAR-Lisflood model is used in catchment mode, the simulations have no representation of suspended sediments and bed rock, and the dune and soil evolution modules are not used. For each catchment, we assess the 15 user-defined parameters against a set of 15 model functions. Finally, we also assess the changes in elevations across different sections of the catchments.

For clarity, we here define some terms used frequently throughout this manuscript:

- Parameter – Adjustable value within a model. The value is determined during model set-up and remains constant throughout a given simulation. The value is often based on recorded values or adjusted during calibration.
- Objective function – an error score between model outputs and observations used to evaluate model performance.
- Model function – a measure derived from model outputs used to evaluate model behaviour in lieu of an adequate objective function.
- Elementary effect (EE) – a value used as part of the Morris Method, indicating the change in function value (objective or model) resulting from a change of parameter value during a single repeat.
- Main effect (ME) – the mean of the elementary effects from all repeats, for a specified parameter and a specified function.

~~The test applies the MM method to perform a global SA on the CAESAR-Lisflood model for two contrasting catchments – the Upper Swale, UK (medium sized, temperate, perennial), and Tin Camp Creek, Australia (small sized, tropical, ephemeral). For each catchment, 15 user-defined parameters are assessed against a set of 15 model functions. Finally, the changes in elevations across the catchments are assessed.~~

2.1 CAESAR-Lisflood

The LEM used is the CAESAR-Lisflood model (Coulthard et al., 2013). CAESAR-Lisflood is a second generation LEM, capable of simulations with greater physical realism than first generation models but also with increased complexity – the model features a large number of fixed, physically-based, or user-defined parameters. ~~This additional complexity may result in an increased non-linearity and sensitivity to model parameters.~~We used CAESAR-Lisflood v1.8, without any additional modifications to the model's functionality from the version freely available online.

A full description of the CAESAR-Lisflood model can be found in Coulthard et al. (2013), and its core functionality is only summarised here. The model utilises an initial DEM built from a regular grid of cells, and in the catchment mode (as used in this model set up) is driven by a rainfall timeseries which can be lumped or spatially distributed (Coulthard and Skinner, 2016b). At each timestep the rainfall input is converted to surface runoff using ~~a version of~~ TOPMODEL (Beven and Kirkby, 1979), and distributed across the catchment and routed using the Lisflood-FP component (Bates et al., 2010). The CAESAR component of the model drives the landscape development using sediment transport formulae based on flow depths and velocities derived from the Lisflood-FP component. Bed load is distributed to neighbouring cells proportionally based on relative bed elevations. This study has not used the suspended sediment processes in the model. The model can handle nine different grain sizes, and information is stored in surface and sub-surface layers where only the top surface layer is 'active' for erosion and deposition. A comprehensive description of this process can be found in (Van De Wiel et al., 2007).~~The Lisflood-FP component generates flow depths and velocities, which are used by the CAESAR component to simulate fluvial erosion, transport and deposition, across 9 grain sizes, using an active layer system, and altering the elevation values of the grid (Van De Wiel et al., 2007).~~

CAESAR-Lisflood is freely available and since 1996 there have been ~~62~~ over 60 published studies using the model over a wide range of temporal and spatial scales (Skinner and Coulthard, 2017). These previous studies provide useful background into model parameter interactions helping to inform the choice of the user-defined parameters used for the SA as described in Section 2.4. Some studies have also investigated the model's sensitivities to external factors - for example, Coulthard and Skinner (2016) investigated the sensitivity of the CAESAR-Lisflood model to the spatial and temporal resolution of precipitation. Other studies have investigated the influence of individual processes or forcings.~~which could be described as both SA and an exploratory test with,~~ F~~f~~or example, Coulthard and Van De Wiel (2017) examined how land-use influences the outputs of the model.~~examining how~~

Formatted: Dutch (Netherlands)

Field Code Changed

Formatted: Dutch (Netherlands)

Field Code Changed

Formatted: Dutch (Netherlands)

Formatted: Dutch (Netherlands)

Formatted: Dutch (Netherlands)

Formatted: Dutch (Netherlands)

Field Code Changed

~~the use of a spatially variable 'm' parameter, representing the land use of an area, could influence the outputs of the model.~~

2.2 Morris Method

Our study used the MM described in Ziliani et al. (2013), i.e. the original MM of Morris (1991), as extended by Campolongo et al. (2007), and applied the "sensitivity" package in the R Statistical Environment (Pujol, 2009) to generate the parameter sets for the SA.

To set up the MM we selected a number of parameters to be assessed, specifying a minimum and maximum range for each, plus a number of iterative steps. The parameter values are equally spaced based on the range and number of steps – for example, a parameter with a range of 2 to 10 and 5 iterative steps would have available values of 2, 4, 6, 8, and 10. This is done for each parameter and, where possible, the same number of iterative steps was used for each.

The MM samples the global parameter space by performing multiple local SAs referred to as repeats. The first simulation in each repeat is made up of a randomly assigned selection of parameter values from the available values. To set up the second simulation in the repeat a single parameter is randomly selected and its value changed by a random number of iterative steps – if we use the example above, if simulation 1 used the value 4, changing this to 2 or 6 would be one iterative step change, to 8 would be two steps, and using 10 would be three steps. For simulation 3 in the repeat another randomly selected parameter is changed although previously changed parameters are no longer available to be selected. This is continued until no further parameters are available to be changed, therefore in our study each repeat contains 16 tests – 1 starting set of parameters, plus 15 parameter changes. In this study we have used 100 repeats, for a total of 1600 individual simulations – for comparison, the implementation of the MM by ~~(~~Ziliani et al.~~,~~ (2013)~~.~~ used 10 repeats.

The sensitivity of the model to changes in parameter values is evaluated by the changes of objective function values between sequential tests within repeats relative to the number of incremental steps the parameter value has been changed by. The change in objective function score between two sequential tests divided by the number of incremental step changes is an elementary effect (EE) of that objective function and the parameter changed (Equation 1). After all 1600 tests have been performed, the main effect (ME) for each objective function and parameter is calculated from the mean of the relevant EEs – the higher the ME the greater the model's sensitivity. Alongside the ME, the standard deviation of the EEs is also calculated as this provides an indication of the non-linearity within the model.

~~Hydrological models faced similar issues to LEMs in the past, in regards to model complexity and resulting processing times when applying SA. To overcome them, hydrologists have used the method of Morris (1991). The MM can be regarded as a global SA, although it actually performs multiple local SA sampled from across the full parameter space – this produces a series of local evaluations, the mean of which is an approximation of the global variance (Saltelli et al., 2000; van Griensven et al., 2006; Norton, 2009). The main strength of the MM is its computational efficiency. Herman et al. (2013) showed that the MM could estimate similar variance in model outputs to the Sobol' Variance-based global SA method (Sobol', 2001), yet required 300 times less evaluations, and significant less data storage for an application to a distributed catchment hydrological model. The robustness of this approach has been further shown by numerous workers (e.g. Brockmann and Morgenroth, 2007; Pappenberger et al., 2008; Yang, 2011). However, the MM cannot provide a full quantitative assessment of parameter sensitivity and is dependent upon the user-defined bounds to the parameter space. It can successfully rank parameters between the least and most influential to model outputs, but cannot determine parameters' exact relative influence (Brockmann and Morgenroth, 2007). This combination of advantage and limitation has seen it used extensively as a pre-screening stage,~~

~~isolating the most influential parameters for further SA with quantitative, yet more computationally expensive, methods (e.g. Ratto et al., 2007; Song et al., 2015; Yang, 2011; Ziliani et al., 2013).~~

~~This paper uses the MM described in Ziliani et al. (2013), i.e. the original MM of Morris (1991), as extended by Campolongo et al. (2007). The Design of Experiment (DOE) used the R Statistical Environment and the "sensitivity" software package (Pujol, 2009). A full description of the method can be found in these papers, but a summary is provided here.~~

~~MM operates as a series of local SA starting with a stochastically selected set of initial parameters drawn from the global parameter space. For each parameter, a defined boundary of available values is set and divided into a series of equal incremental steps - each parameter is subsequently altered one-at-a-time (OAT) by a number of incremental steps, until each parameter has been altered once. The change in the model output measures, or model factors, is observed between each test – these are the Elementary Effects (EE). This process is repeated a number of times and the mean and standard deviations of the EE for each parameter is calculated – the Main Effect (ME). The higher the mean ME for a parameter shows greater influence of that parameter on the factor, and a higher standard deviation indicates greater non-linear relationships with other parameters. The calculation of an EE can be summarised as in Equation 1 –~~

Equation 1

$$d_{ij} = \left| \frac{y(x_1 x_2 \ldots, x_{i-1}, x_i + \Delta_i, x_{i+1}, \ldots, x_k) - y\,(x_1 x_2 \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k)}{\Delta_i} \right|$$

where d_{ij} is the value of the j^{th} EE ($j = 1, \ldots, r$; where *r* is the number of repetitions (here *r* = 100)) of the i^{th} parameter (e.g. *i* =1 refers to sediment transport formula, see Table 1), x_i is the value of the i^{th} parameter, *k* is the number of parameters investigated (here 15), $y(x_1, x_2, \ldots, x_k)$ is the value of the selected objective function, and Δ_i is the change in incremental steps parameter *i* was altered by.

Formatted: Font: Bold

~~where, d_{ij} is the *j*th EE of the *i*th model output measure (eg, *i* =1 refers to Sediment Transport Formula, see Table 1), *k* is the number of parameters investigated (here 15), $y(x_1 x_{2,...,} x_k)$ is the value of the model output measure, *r* is the number of repetitions (here r = 100), and Δ_i is the change in incremental steps parameter *i* was altered by.~~

2.3 Study Basins

2.3.1. Upper Swale, UK

The Swale catchment, UK, is a medium sized basin (181 km^2) with 500 m of ~~elevation drop~~relief (Figure 1). It has been used extensively in previous CAESAR/CAESAR-Lisflood applications (Coulthard et al., 2012; Coulthard and Macklin, 2001; Coulthard and Skinner, 2016a; Coulthard and Van De Wiel, 2013). For this SA, it represents a medium basin in a temperate climate. All simulations on the Swale are ~~based on~~use a 50 m resolution DEM based on airborne LiDAR.~~and 30 years in duration.~~ Precipitation inputs are 10 years of NIMROD composite RADAR rainfall estimates (Met Office, 2003), applied at a 1 h temporal and 5 km spatial resolution, and repeated three times for a 30 year timeseries.~~(1 h – 5 km resolution) repeated.~~

2.3.2. Tin Camp Creek, Australia

The Tin Camp Creek catchment is a small sub-catchment (0.5 km^2) of the full Tin Camp Creek system (Hancock et al., 2010; Hancock, 2006) (Figure 1). The basin has ~~a~~ 45 m of ~~elevation drop~~relief and is in the tropical region of the Northern Territory, Australia. ~~Contrasting~~ In contrast to the Swale, Tin Camp Creek is ~~much smaller~~a small basin and the region has pronounced wet and dry seasons, with short intense rainstorms a feature of wet season precipitation. The DEM is at 10 m grid cell resolution produced from high resolution digital photogrammetry (Hancock, 2012).~~, and like the Swale~~

~~simulations are 30 years in length.~~ The rainfall input is taken from observations from a single raingauge at Jabiru Airport, providing a 1 h – lumped (single catchment-average) resolution timeseries for 23 years, with the first 7 years repeated to produce a continuous 30 year timeseries.~~which was looped to create the 30 year record required~~.

2.3.2 Stream Orders

Formatted: Font: Bold

~~For both basins the~~The changes in the mean elevation across different areas of the catchments ~~will be~~were assessed as ~~a representation~~an illustration of spatial differences in geomorphic change. ~~of changes in the geomorphology.~~ Each basin was sub-divided into regions corresponding to the watersheds of five stream orders based on the proportion of the catchment drained in the initial DEM – 1^{st} = < 1 %; 2^{nd} = > 1 %; 3^{rd} = > 10 %; 4^{th} = > 25 %; 5^{th} = > 50 % (see Figure 1). This method is novel and was developed to provide a consistent method of sub-dividing both catchments independent of factors such as connectivity and DEM resolution.

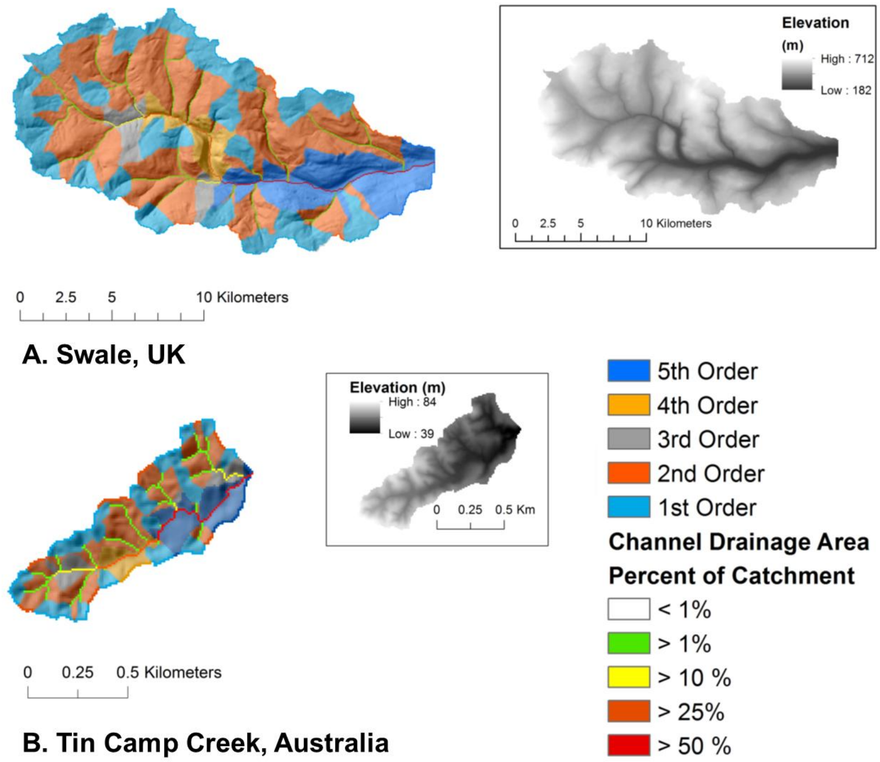


Figure 1 – Elevation map for the Upper Swale catchment, UK (top), and Tin Camp Creek catchment, Australia (bottom) – note the differences in scale between catchments. Each catchment is sub-divided into watersheds of five stream orders based on the proportion of the catchment drained.

2.4 User-Defined Parameters

The MM implemented here used 15 user-defined parameters, each with 5 iterative step values (as described in Section 2.2). The only exception was the choice of sediment transport formula parameter

Formatted: Justified

(SED, Table 1) where only two options are available. The parameters, their ranges, and available values are shown in Table 1.

Formatted: Font: 11 pt, Not Bold

Table 1 – User-defined parameters used and the min-max values for the two study catchments.

Code	Parameter	Steps	Upper Swale	Tin Camp Creek
(1) SED	Sediment Transport Formula	2	1 Wilcock & Crowe / 2 Einstein	1 Wilcock & Crowe / 2 Einstein
(2) MEL	Max Erode Limit (m)	5	0.01; 0.015; 0.02; 0.025; 0.03	0.001; 0.0015; 0.002; 0.0025; 0.003
(3) CLR	In Channel Lateral Erosion Rate	5	10; 15; 20; 25; 30	10; 15; 20; 25; 30
(4) LAT	Lateral Erosion Rate	5	$2.5e^{-6}$; $3.75e^{-6}$; $5e^{-6}$; $6.25e^{-6}$; $7.5e^{-6}$	$1.5e^{-6}$; $2.25e^{-6}$; $3e^{-6}$; $3.75e^{-6}$; $4.5e^{-6}$
(5) VEG	Vegetation Critical Shear Stress (Pa)	5	10; 15; 20; 25; 30	2; 3.25; 4.5; 5.75; 7
(6) MAT	Grass Maturity Rate (yr)	5	0.5; 0.75; 1; 1.25; 1.5	0.5; 0.875; 1.25; 1.625; 2
(7) SCR	Soil Creep Rate (m/yr)	5	0.00125; 0.001875; 0.0025; 0.003125; 0.00375	0.00125; 0.001875; 0.0025; 0.003125; 0.00375
(8) SFT	Slope Failure Threshold (°)	5	40; 42.5; 45; 47.5; 50	40; 42.5; 45; 47.5; 50
(9) IOD	In/Out Difference ($m^3.s^{-1}$)	5	2.5; 3.75; 5; 6.25; 7.5	0.1; 0.175; 0.25; 0.325; 0.4
(10) MinQ	Min Q Value (m)	5	0.25; 0.375; 0.5; 0.625; 0.75	0.025; 0.0375; 0.05; 0.0625; 0.075
(11) MaxQ	Max Q Value (m)	5	2.5; 3.75; 5; 6.25; 7.5	2.5; 3.75; 5; 6.25; 7.5
(12) SEC	Slope for Edge Cells	5	0.0025; 0.00375; 0.005; 0.00625; 0.0075	0.0025; 0.00375; 0.005; 0.00625; 0.0075
(13) EVR	Evaporation Rate (m/d)	5	0.00067; 0.001005; 0.00134; 0.001675; 0.00201	0.0025; 0.004375; 0.00625; 0.008125; 0.01
(14) MNR	Manning's n Roughness	5	0.03; 0.035; 0.04; 0.045; 0.05	0.03; 0.0325; 0.035; 0.0375; 0.04
(15) GSS	Grain Size Set	5	Set 1; Set 2; Set 3; Set 4; Set 5	Set 1; Set 2; Set 3; Set 4; Set 5

The MM varies the value of each parameter tested once per repeat, and here we use 100 repeats. Therefore, careful consideration was required in the selection of parameters as each parameter tested added 100 model runs to the test – there are 49 user-defined parameters in the version of CAESAR-Lisflood model used (v1.8), and even excluding parameters associated with dune and soil development, there are still 35 user-defined parameters. To test each would require 3600 model runs for each catchment, yet the inclusion of some parameters is likely to add little value. ~~Therefore, in total,~~Thus this was narrowed to a set of 15 user-defined parameters ~~were tested~~ (Table 1) ~~and~~ with the selection ~~was~~ based largely on prior knowledge of the importance of these parameters, or due to a lack of previous knowledge of the influence of the parameters on the model – full justification of the selection of parameters, and descriptions of their purpose within the model, can be found in ~~Supplement S1 of the~~ Supplementary Material S1.

The MM is subjective in that the relative sensitivities shown depend on the minimum and maximum range values set by the user. Therefore, it is necessary to set each parameter's range to be broadly equal to the others in order to obtain useful information. To be consistent, where possible we have used a default value taken from past calibrations and varied this by +/- 25 % and +/- 50 %. There are some instances where this was not appropriate and a minimum and maximum bound was set instead, with 5 iterative steps of equal distance determined (for example, the Manning's n Roughness for Tin Camp Creek where +/- 50 % would have resulted in obviously physically unrealistic values – see Table 1 for values used).

The sediment transport formulae employed for SED were Einstein (Einstein, 1950) and Wilcock & Crowe (Wilcock and Crowe, 2003). These were not selected as representing the best fit for the catchments simulated but because they are the formulae available in the unmodified version of CAESAR-Lisflood. The sediment transport formulae parameter was applied as a binary choice, with the model switching from one formula to the other once per repeat (no other parameter values were varied when this occurs, as per the description of the MM in Section 2.2). It was assumed that this change constituted a single iterative step change for calculating related EEs.

~~The Morris Method is a qualitative method and the results are subjective on the range of values and number of iterative steps set by the user. Therefore, it is necessary to set each parameter's range to be broadly equal to the others. Whilst it is difficult to define what this means it is also difficult to~~

~~estimate without prior knowledge - something this study is attempting to address. Here, as a general rule, we have used a default value taken from past calibrations and varied this by +/- 25 % and +/- 50 %. There are some instances where this was not appropriate, and instead a minimum and maximum bound was set instead, and 5 iterative steps of equal distance determined (for example, the Manning's n Roughness for Tin Camp Creek).~~

~~The Sediment Transport Laws employed for SED were Einstein (Einstein, 1951) and Wilcock & Crowe (Wilcock and Crowe, 2003). This was applied as a binary two-step, switching from one Law to the other.~~

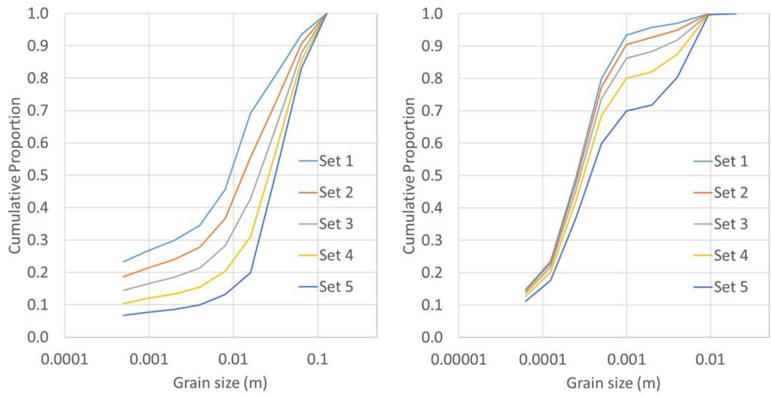


Figure 2 – Sediment grain size distribution sets for the Upper Swale (left) and Tin Camp Creek (right), showing the cumulative proportions.

Grain size distribution has been shown to influence erosion patterns and erosion rate(Hancock and Coulthard, 2012)). It is more difficult to define iterative steps for the sediment grain size sets which include 9 different grain sizes and proportions in each. Instead, these were skewed by altering the proportions of the five smallest grain sizes +/- 25 % and 50 %, and the opposite to the four largest grain sizes, before adjusting the final proportions to equal one based on the relative values. This produces two sets biased for smaller grain sizes (Sets 1 and 2), and two sets biased for larger grain sizes (Sets 4 and 5), as well as the default grain size set (Set 3) (Figure2).~~. The grain size distributions can be seen in Figure 2.~~ Note, that the grain size sets presented in Figure 2 contain non-cohesive silts and this requires an extrapolation of the sediment transport formulae (Van De Wiel et al., 2007).

2.5 Model Functions

The common method of assessing a model's sensitivity to parameters values via SA, and the method employed by the MM, is to observe the variations to objective function measures. However, the difficulties in applying an objective function approach to LEMs were highlighted in Section 1.2, and in order to apply an SA a novel approach is required. The method we have developed eschews the objective function approach and instead assesses the model against a series of model functions designed to reflect some of the core behaviours displayed in the model – these can be seen in Table 2. This represents a philosophical difference to traditional applications of SA – here we are not testing the model against its skill in simulating the physical environment, but rather how the model responds behaviourally to changes in the user-defined parameters detailed in Section 2.4. The 15 model functions (Table 2) are simple, scalable and transferable between different catchment types, and can be applied to simulations of different timeframes. The model functions are based on outputs which are not unique to CAESAR-Lisflood, so can be applied to other LEM and geomorphic models.~~The common method of assessing a model's sensitivity to parameters values via SA is to observe the variations to objective function measures, yet as discussed in Section 1.3 the use of objective functions is often not feasible or appropriate when simulating using LEMs. Also in Section 1.3, previous attempts to quantify changes to the geomorphology of catchments were discussed, showing that no statistical methods, whether based on catchment outlet or some feature of the landscape within the catchment,~~

~~fully captured or reflected the geomorphic change. The methods reviewed in Hancock and Willgoose (2001) have also been shown to be of little value for simulations of 1000 years and less.~~

Table 2 – Model Functions and the associated core behaviours.

Model Function	Core Behaviour
Total Sediment Yield (m^3)	Catchment Sediment Yield
Mean Daily Sediment Yield (m^3)	
Peak Daily Sediment Yield (m^3)	
Time to Peak Sediment Yield (s)	
Days when Sediment Yield > Baseline (d)	
Total Net Erosion (m^3)	Internal Geomorphology
Total Net Deposition (m^3)	
Area with > 0.02 m Erosion (m^2)	
Area with > 0.02 m Deposition (m^2)	
Total Discharge (m^3)	Catchment Discharge
Mean Daily Discharge (m^3)	
Peak Daily Discharge (m^3)	
Time to Peak Discharge (s)	
Days when Discharge > Baseline (d)	
Total Model Iterations (calculations)	Model Efficiency

The model functions were applied to the MM as described in Section 2.2, substituting the model functions in place of the objective functions with no further changes to the method. Model function values were calculated at the end of each simulation.

~~The method we have used is to abandon the objective function approach and instead assess the model against a series of Model Functions designed to reflect some of the core behaviours displayed in the model. It should be noted that this is a philosophical difference to traditional applications of SA – here we are not testing the model against its skill in simulating the physical environment, but rather how the model responds behaviourally to the uncertainty in the user-defined parameters detailed in Section 2.4 – in this sense it also differs from methods of assessing parameter uncertainty, such as the Generalised Likelihood Uncertainty Estimation (GLUE) of Beven and Binley (1992), yet is an important step towards the adoption of such techniques with LEMs. The 15 model functions (Table 2) are simple, scalable and transferable between different catchment types, and can be applied to simulations of~~

~~different timeframes. The model functions are based on outputs which are not unique to CAESAR-Lisflood, so can be applied to other LEM and geomorphic models.~~

~~The~~ To summarise the large amount of information produced, the ME of each parameter ~~versus each~~and model function combination was normalised based on the proportion of the ME for highest ranking parameter for that model function – therefore the highest ranked parameter for each model function always scored 1. The scores for each parameter were aggregated for across all model functions based on the mean of the scores. The model functions were sub-divided into core behaviour groups (Table 2), and the scores aggregated again for each core behaviour. The same was also done, separately, for the standard deviations of each parameter and model function.

~~LEMs are subject to transient model behaviour (an internal model adjustment), as the model reacts to effects of the initial DEM surface and the global grain size distribution. During the initial period of model simulation this results in accelerated sediment processes as the model removes uneven surfaces and noise, and easily mobilises smaller grain sizes in the channel. This is commonly accounted for by allowing the model to run for a 'spin-up' period before the simulation begins. It is possible that small differences in the model could be exaggerated during this period, therefore the first 10 years of each simulation has been discounted for the calculation of the model functions.~~

3. Results

3.1 All Model Functions

Figure 3 shows the spread of parameter influence for both catchments, where ~~the~~ a higher ~~the~~ mean of the aggregated MEs indicates greater sensitivity in the model to that parameter, and ~~the~~ ahigher standard deviation shows greater non-linearity ~~to~~ when interacting with other parameters. Table 3 shows the parameters ranked for both catchments, based on the aggregated mean ME values. The most influential parameter is SED (see Table 1 for full description of parameter abreviations), ranked top for both catchments and also being most influential by a reasonable margin, having an aggregated mean of at least 0.2 higher than the 2nd ranked parameter. Other parameters, such as VEG, IOD, MNR, MinQ and GSS, rank highly or mid-range. There is a visually close correlation between the most influential parameters and those ~~which~~ that display the most non-linearity (Figure 3).

Table 3 – Parameters ranked by means for each catchment from the aggregated scores for all Elementary Effects. SED = sediment transport formula; MEL = maximum erode limit; CLR = in channel lateral erosion rate; LAT = lateral erosion rate; VEG = vegetation critical shear stress; MAT = grass maturity rate; SCR = soil creep rate; SFT = slope failure threshold; IOD in/out difference; MinQ = minimum Q value; MaxQ = maximum Q value; SEC = slope for edge cells; EVR = evaporation rate; MNR = Manning's n roughness coefficient; and GSS = grain size set.

Rank (by mean: 1 = most influential)	Upper Swale	Tin Camp Creek
1	SED	SED
2	MNR	SEC
3	IOD	VEG
4	GSS	GSS
5	EVR	MinQ
6	VEG	IOD
7	MinQ	MNR
8	LAT	MAT
9	CLR	SCR
10	SCR	MEL
11	SEC	LAT

12	MAT	CLR
13	MEL	MaxQ
14	MaxQ	SFT
15	SFT	EVR

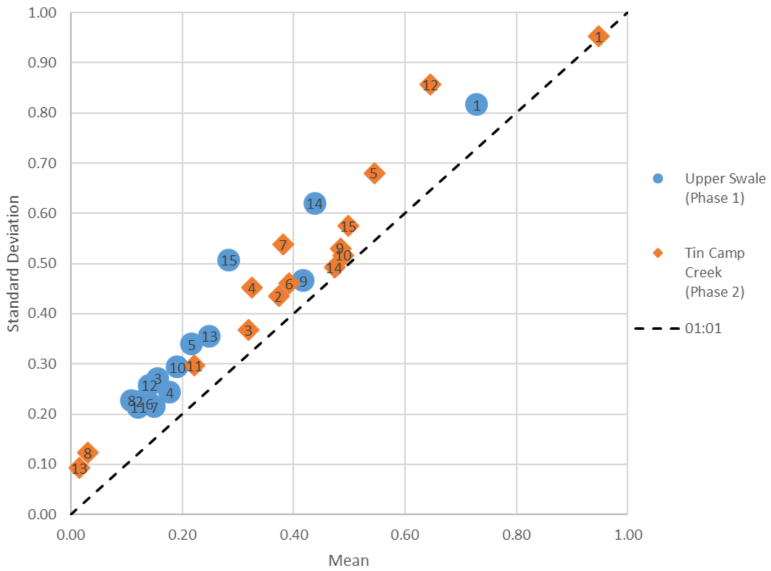


Figure 3 – Aggregated scores for all Elementary Effects where: 1 = sediment transport formula (SED); 2 = maximum erode limit (MEL); 3 = in channel lateral erosion rate (CLR); 4 = lateral erosion rate (LAT); 5 = critical vegetation shear stress (VEG); 6 = grass maturity rate (MAT); 7 = soil creep rate (SCR); 8 = slope failure threshold (SFT); 9 = in/out difference (IOD); 10 = minimum Q value (MinQ); 11 = maximum Q value (MaxQ); 12 = slope for edge cells (SEC); 13 = evaporation rate (EVR); 14 = Manning's n roughness coefficient (MNR); and 15 = grain size set (GSS).~~1 = SED; 2 = MEL; 3 = CLR; 4 = LAT; 5 = VEG; 6 = MAT; 7 = SCR; 8 = SFT; 9 = IOD; 10 = MinQ; 11 = MaxQ; 12 = SEC; 13 = EVR; 14 = MNR; and 15 = GSS.~~

3.2 Catchment Sediment Yield Vs Internal Geomorphology

The core behaviours of c~~C~~atchment s~~S~~ediment y~~Y~~ield and i~~I~~nternal g~~G~~eomorphology show a different response to the changes in parameter values, as can be seen in Figure 4, and also the rankings in Table 4. For both catchments, SED is ranked as most influential for c~~C~~atchment s~~S~~ediment y~~Y~~ields. For influence on the i~~I~~nternal g~~G~~eomorphology, SEC ranks higher in the Tin Camp Creek catchment. The Upper Swale catchment displays a similar response with both behaviours, with SED and MNR most influential and by similar amounts, although GSS has less influence on i~~I~~nternal g~~G~~eomorphology. The change in response for Tin Camp Creek is more varied – SED is less influential on i~~I~~nternal g~~G~~eomorphology, and SEC is the most influential with a higher aggregated mean. GSS is slightly less influential, and MNR slightly more, and VEG is more influential on the i~~I~~nternal g~~G~~eomorphology than it is on c~~C~~atchment s~~S~~ediment y~~Y~~ield. For both model functions, there again is a strong visually correlation between those parameters showing the most influence and those showing the most non-linear behaviour.

Table 4 – Parameters ranked by means for each catchment from the aggregated scores for catchment sediment yields (SY) and internal geomorphology (IG) elementary effects. SED = sediment transport formula; MEL = maximum erode limit; CLR = in channel lateral erosion rate; LAT = lateral erosion rate; VEG = vegetation critical shear stress; MAT = grass maturity rate; SCR = soil creep rate; SFT = slope failure threshold; IOD in/out difference; MinQ = minimum Q value; MaxQ = maximum Q value; SEC = slope for edge cells; EVR = evaporation rate; MNR = Manning's n roughness coefficient; and GSS = grain size set.
~~**Parameters ranked by means for each catchment from the aggregated scores for Catchment Sediment Yields (SY) and Internal Geomorphology Elementary Effects (IG).**~~

Rank (by mean: 1 = most influential)	Upper Swale		Tin Camp Creek	
	SY	IG	SY	IG
1	SED	SED	SED	SEC
2	MNR	MNR	SEC	SED

3	GSS	GSS	GSS	VEG
4	LAT	VEG	MinQ	MNR
5	VEG	CLR	VEG	MinQ
6	EVR	LAT	MNR	GSS
7	MinQ	MinQ	IOD	SCR
8	SCR	MaxQ	MAT	MAT
9	IOD	EVR	SCR	IOD
10	SEC	IOD	MEL	LAT
11	MAT	MAT	CLR	MEL
12	SFT	SEC	LAT	CLR
13	CLR	SCR	MaxQ	MaxQ
14	MEL	MEL	SFT	SFT
15	MaxQ	SFT	EVR	EVR

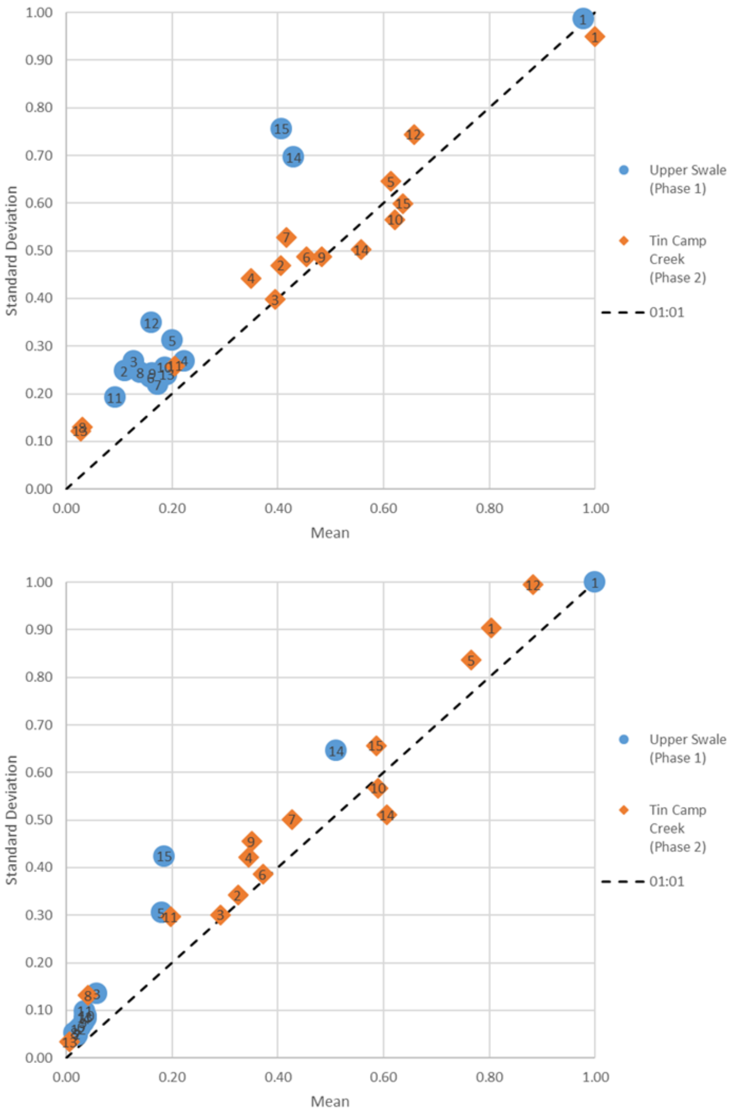
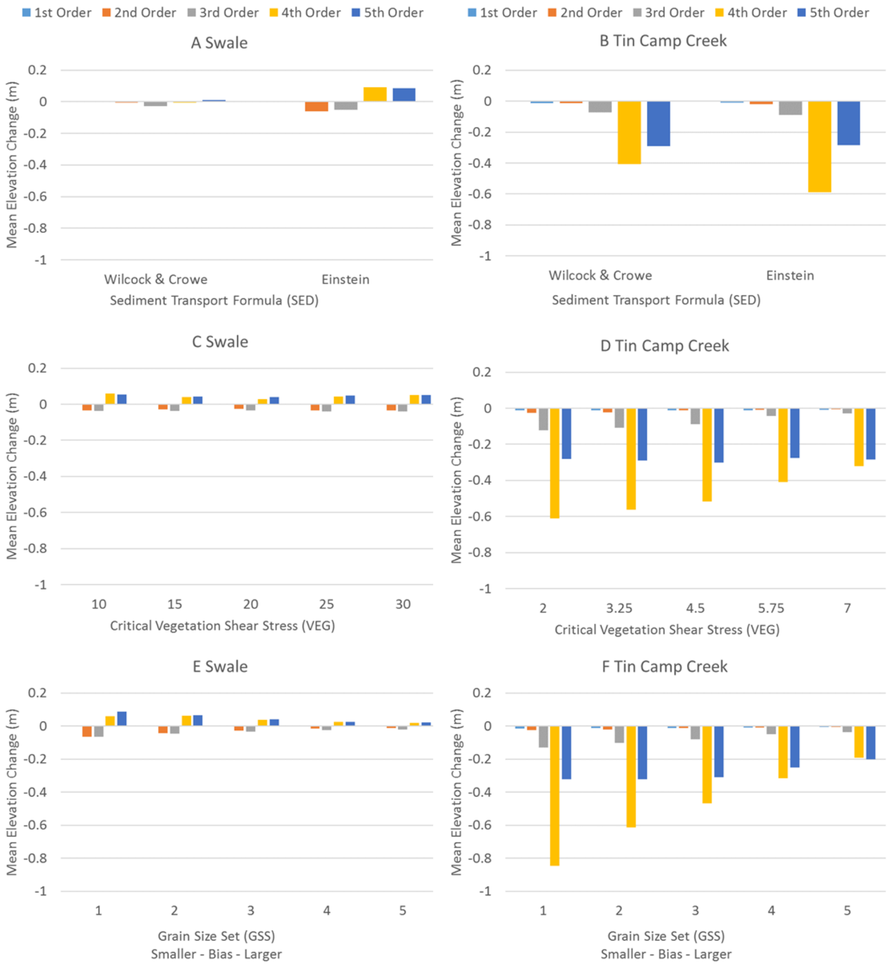


Figure 4 – Aggregated scores for sediment yield (top) and internal geomorphology (bottom) where: 1 = sediment transport formula (SED); 2 = maximum erode limit (MEL); 3 = in channel lateral erosion rate (CLR); 4 = lateral erosion rate (LAT); 5 = critical vegetation shear stress (VEG); 6 = grass maturity rate (MAT); 7 = soil creep rate (SCR); 8 = slope failure threshold (SFT); 9 = in/out difference (IOD); 10 = minimum Q value (MinQ);

11 = maximum Q value (MaxQ); 12 = slope for edge cells (SEC); 13 = evaporation rate (EVR); 14 = Manning's n roughness coefficient (MNR); and 15 = grain size set (GSS).~~Aggregated scores for Sediment Yield Elementary Effects (top) and Internal Geomorphology (bottom) where: 1 = SED; 2 = MEL; 3 = CLR; 4 = LAT; 5 = VEG; 6 = MAT; 7 = SCR; 8 = SFT; 9 = IOD; 10 = MinQ; 11 = MaxQ; 12 = SEC; 13 = EVR; 14 = MNR; and 15 = GSS.~~

3.3 Changes in the Mean Elevations


1st Order 2nd Order 3rd Order 4th Order 5th Order
A Swale
B Tin Camp Creek
Mean Elevation Change (m)
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Wilcock & Crowe
Einstein
Sediment Transport Formula (SED)
C Swale
D Tin Camp Creek
10 15 20 25 30
2 3.25 4.5 5.75 7
Critical Vegetation Shear Stress (VEG)
E Swale
F Tin Camp Creek
1 2 3 4 5
Grain Size Set (GSS)
Smaller - Bias - Larger

642

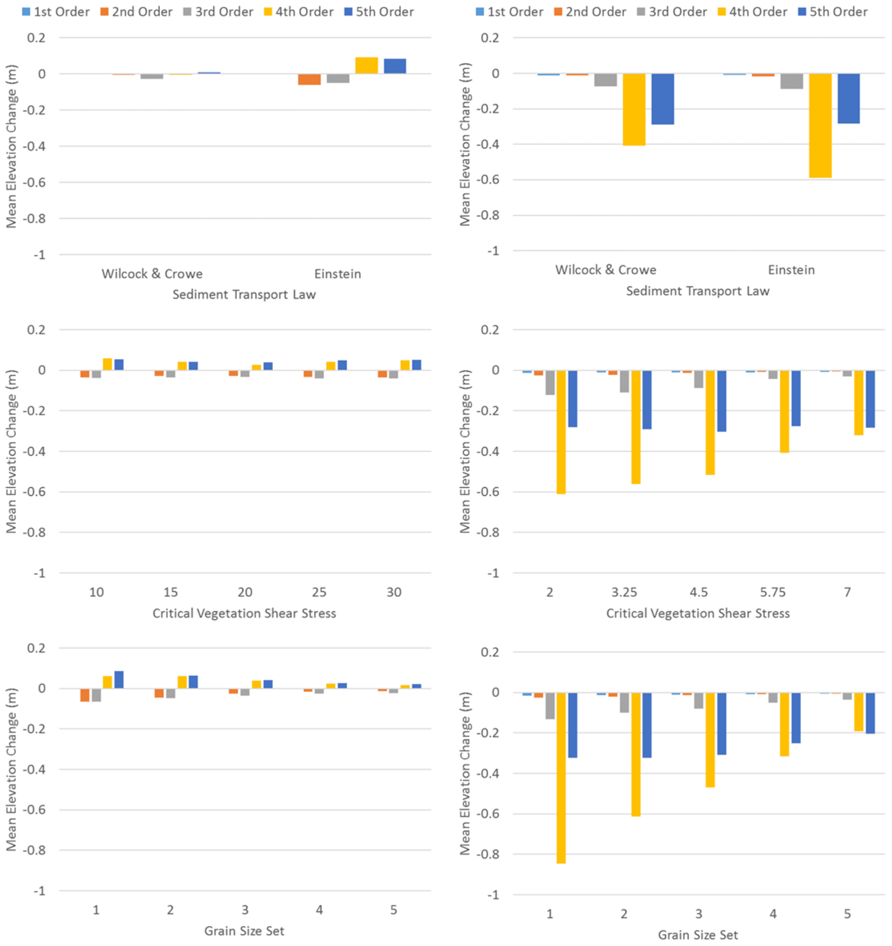


Figure 5 – Illustration of changes in the mean elevations for Upper Swale (A, C and E), and Tin Camp Creek (B, D and F) for the tests split by SED (A and B), VEG (C and D), and GSS (E and F) where 1 and 2 are biased smaller, and 4 and 5 are biased larger. The catchment is sub-divided into watersheds of five stream orders, based on proportion of catchment drained.~~Changes in the mean elevations for Upper Swale (left), and Tin Camp Creek (right) for the tests split by SED (top), VEG (middle), and GSS (bottom). The catchment is sub-divided into watersheds of five stream orders, based on proportion of catchment drained.~~

The test results were binned by the parameter values used, and the mean changes in the mean elevations across the 5 stream orders calculated – Figure 5 illustrates how changes in parameter values might influence the spatial patterns of landscape change using SED, VEG and GSS as examples. For SED (Fig 5.A and 5.B), the most obvious difference is the scale of changes seen using each formula with Einstein generally showing greater change. For Tin Camp Creek (Fig 5.B) the spatial changes are similar, but for the larger Swale (Fig 5.A) there are differences in relative rates in 2nd and 4th order areas. In the Swale, VEG (Fig 5.C) appears to have little impact on the patterns and scale of changes, yet in Tin Camp Creek (Fig 5.D) there is reduction in the rates of erosion across the catchment with higher values, except in the 5th order areas which remain at a similar level. Finally, both catchments show a reduction in rates of erosion with a greater proportion of larger grain sizes, yet this is more pronounced 4th order areas in Tin Camp Creek (Fig 5.F).~~shows the changes in each catchment for parameters SED, VEG and GSS. In general, the patterns of changes remain similar despite changing parameter values, yet rates of change do vary – for example, for GSS, the mean reduction in elevations decreases across the catchments using grain size sets biased towards larger grain sizes. In both catchments, the largest variations are observed in the 4th and 5th stream orders.~~

4. Discussion

The results reveal some important insights into the application of the SA to LEMs generally, and also on specific behaviours of the CAESAR-Lisflood model. Here we discuss model functions (Section 4.1), sediment transport formulae (Section 4.2), implications for calibrating LEMs (Section 4.3), full uncertainty analyses of LEMs (Section 4.4), and limitations of this study (Section 4.5).

~~The SA has been applied here to a single LEM, CAESAR-Lisflood, and the implications for that model have been discussed above. Yet the results also reveal some important insights concerning metrics, transferability, sediment transport laws, and full uncertainty analysis, which are relevant to all LEMS.~~

~~1. Metrics~~4.1 Model Functions

Formatted: Font: Bold

~~Interestingly, the~~Our findings show that different ~~metrics~~ model functions provide us with different indications of model sensitivity. This has important implications for how to measure LEM performance – and more widely how to quantify and assess geomorphic change within a basin. For example, Figure 4 and Table 4 show how any LEM assessment must depend on the applied metric for comparison. Model functions that quantify sediment yield (derived at the catchment outlet) indicate different sensitivities compared to model functions that quantify the internal landform response ~~that the model has different responses when assessed using sediment yield model functions (calculated from the catchment outlet) to when using the internal geomorphology model functions~~ (based on spatial measures from within the catchment). Whilst at-a-point sediment yields are straightforward to extract from simulation data and easily related to field measurements (e.g. gauges, although these have their own associated uncertainties), similar or identical yields may conceal very different behaviours within the basin. This highlights an important aspect of LEM calibration:~~is important for users to realise that when calibrating LEMs,~~ changes in sediment yields from a catchment outlet only provide partial information of what is changing internally. We therefore argue that metrics incorporating *spatial* changes in the basin (as well as bulk figures) are vital for assessing LEM performance. (i.e. time series of high resolution DEM data from LiDAR/photogrammetry~~a nested set of flumes within a catchment to quantify discharge and sediment output~~) This is especially important as the shape of the landscape – where material has been eroded and deposited – is effectively the basins geomorphic memory and will directly influence subsequent model performance. For other basin scale models (e.g. hydrological models) this aspect is possibly not so important over longer-terms given the limited temporal extent memory of basin antecedence. Some of the challenges of LEM output comparison are similar to those of meteorology/climatology and may require a shift in expectation from end users as to what is possible. For example, predicting detailed patterns of local erosion and deposition is akin to predicting weather (low comparability especially over longer time scales) but more general (spatial and temporal) patterns of basin change are similar to climate predictions (better comparability especially for longer time scales).

~~*2. Transferability*~~

~~For environmental models, a single selection of calibrated parameter values is not transferable between catchments as the conditions are different. The same is true for SAs and here we have clear different behaviours between the two catchments tested – some of this can be attributed to the different conditions in each catchment and associated data, but also to the choice of parameter values used in the SA (ie, the minimum and maximum bounds set). The bounds of the parameter values are chosen to be appropriate to the catchment they are applied to. Hence, SA are not transferable between catchments, and should be performed as a preliminary phase for any new investigation. Another consideration is that a single calibrated parameter set is also likely to be non-stationary, especially when factors such as climate and land-use are also non-stationary, and similarly this may impact on model behaviour over time.~~

~~3~~4.2. Sediment Transport Formulae

Formatted: Font: Bold

Our SA shows that the choice of sediment transport formula (SED) had a very strong impact on the model functions.~~,~~ ~~and as~~As sediment transport formulae are also integrated into other LEMs and geomorphic models they will affect their outcomes too. ~~This is, however, to be expected as previous studies have shown that erosion thresholds in sediment transport for LEMs have a significant impact on a model's sensitivity to climate forcings (Tucker, 2004).~~ Looking at sediment transport formulae themselves, ~~(~~Gomez and Church~~,~~ (1989) tested 11 different sediment transport formulae to the same data sets and showed widespread variation in predictions – in some cases over orders of magnitude. The variation in the model performance can be explained by the derivation of the sediment transport formulae themselves, that are often ~~empirically~~ theory-based ~~on~~ but fitted to limited laboratory and field data, sometimes representing temporal averages over equilibrium conditions (Gomez and Church, 1989). The formulae do not, and were likely never intended to, represent the full variation of actual flow conditions in natural river. As LEMs commonly amalgamate a set of geomorphic models or transport formulae, their performance hinges in the a number of individual model components. Therefore, when applied to different situations, they ~~can be wrong~~may not be appropriate~~.~~ (Coulthard et al., 2007a).

4.3 Implications for Calibrating LEMs

Formatted: Font: Bold

This, however, presents ~~researchers using LEMs with a considerable problem~~a challenge, as it is highly likely that the sediment transport formula to be used was ~~not~~ neither designed nor calibrated for ~~that~~ a particular model application. The SIBERIA model (Hancock et al., 2010, 2016, 2017; Hancock and Willgoose, 2001; Willgoose et al., 2003) overcomes this issue by having a version of the Einstein sediment transport ~~law~~ formula (Einstein, 1950) that is calibrated or tuned to field data on erosion rates. However, even when calibrated, LEMs (and their sediment transport formulae) face another hurdle with the non-stationarity of basin sediment yields. For example, a calibrated LEM will be adjusted to perform for a set of observed sediment outputs or erosion and deposition patterns. If, due to climate change for example, rainfall and channel flows significantly increase then the initial calibration may no longer be valid (Coulthard et al., 2007b). This is similar to issues faced by calibrating hydrological models (e.g., ~~(~~Li et al., 2012) though the non-linear sediment response of LEMs like CAESAR-Lisflood (Coulthard et al., 2012) may make LEMs more sensitive to this. Furthermore, this analysis suggests that detailed justification and calibration of model choices around sediment transport will lead to the most effective gains in model skill.~~The issue of non-stationarity has been a considerable focus of the hydrological community in recent years. However, despite all the above limitations, LEMs – when applied correctly – have generally been found to compare well with available~~

~~field data. Nonetheless, the issue of the scaling of parameters for different catchments and even more importantly DEM grid size is an issue that remains to be addressed.~~

4.4 Full Uncertainty Analysis

Formatted: Font: Bold

It is important to note that the MM does not provide an absolute value of sensitivity, but ranks each factor based on its relative influence on the model. This means it can be used to assess the main sources of uncertainty on a particular model set up. The next step ~~would be~~is then to establish how the uncertainty caused by model parameters (e.g. the choice of sediment transport formula) compares to other identified sources of uncertainty, such as rainfall input uncertainty, DEM observation and resolution uncertainty, and length of spin-up period. For example, it may be that the choice of sediment transport formula may only be a minor source of uncertainty compared to the DEM resolution, or equally, it might be the most significant source of uncertainty in a LEM's ouput.

Importantly, whilst the simulation of long-term development of landscapes may be somewhat resilient to some uncertainties, e.g. initial conditions~~y~~ (Hancock et al., 2016), any attempt to reproduce, predict or forecast physical changes~~, especially if there is a decision-making element,~~ should have the same appreciation of uncertainty and rigorous testing that is applied to models in other fields (e.g., hydrology and hydraulics). There are many methods available, but when discussing CAESAR-Lisflood the applications applied to Lisflood-FP seem a reasonable place to start. ~~has been applied to models such as Lisflood-FP. For example, the~~ Lisflood-FP has been rigorously tested and benchmarked for decision-making purposes (Hunter et al., 2005; Neelz & Pender, 2013), and the use of SA to assess model response and uncertainty is standard practise (Di Baldassarre et al., 2009; Fewtrell et al., 2008, 2011; Hall et al., 2005; Horritt and Bates, 2001, 2002; Hunter et al., 2008; Neal et al., 2011; Sampson et al., 2012), often as a stage of calibration using the GLUE method (Aronica et al., 2002; Bates et al., 2004; Horritt et al., 2006; Hunter et al., 2005; Pappenberger et al., 2007; Wong et

Field Code Changed

al., 2015). Uncertainty in model predictions can be accounted for by utilising probabilistic measures and uncertainty cascades (for example, Pappenberger et al., 2005; Stephens et al., 2012). This is not considered unique to CAESAR-Lisflood, and any application of an LEM or other geomorphic model for operational, decision-making or forecasting applications should make full consideration of all associated uncertainties.

4.5. Limitations

Formatted: Font: Bold

The main limitation of the MM is the subjectivity in selection of parameter values and ranges. Here, this has been mitigated by consistently selected ranges of +/- 50 % of a default value obtained from previous calibrations (where feasible). An issue emerges with categorical parameters, such as SED, where multiple values cannot be placed in spectrum across a range between minimum and maximum values. The MM has no formal method for dealing with such categorical parameters, so here it has been assumed that switching from one formula to another is a single iterative step change, and this would be the same even with more choices available. This reflects the purpose of the MM, which is to inform about the relative importance of choices of parameter values on the performance/behaviour of the model~~.~~. However, to assess the impact of this single step-change assumption, we performed a further analysis, where it was assumed that switching formula was a change of four iterative steps. This analysis shows that the relative sensitivity of the model to the sediment transport formula choice becomes less important, with other parameters such as Manning’s n Roughness and grain size sets increasing in relative influence (see Supplementary Material S2 for full results of this analysis).

~~There are limitations to the methodology presented here. The MM should not be considered a quantitative assessment of sensitivity – it is designed to be an efficient pre-screening method to isolate key parameters for further assessment or for calibration, and ranks parameter values based only on their relative influence of the model. It is also subjective in the sense that the user defines the parameter space explored by setting minimum and maximum values. The range of these values and~~

~~the number of iterative steps between them will have an influence on the relative influence shown – here, the fact that SED was binary, with no intermediate steps, whereas most other parameters had five equal and iterative steps, will have affected its overall relative influence. Reducing the number of iterative steps would likely increase the EEs calculated, and increasing would reduce them, and shift the other parameters' relative influence against that for SED. This is acknowledged here, but the range of parameter values and the steps used were appropriate to represent the possible uncertainties in the model (i.e., they were based on proportional deviations from previous calibrated parameter sets).~~

An obvious limitation to this exercise is computational resource. This study~~test~~ incorporated 1600 individual model runs to test the behavioural response of the model to 15 parameters, in just two catchments, and this partly influenced the choice to limit ~~the~~ simulation periods to 2~~3~~0 years. The bulk of simulations used Intel i7-5960X processors and using Solid State Drives (SSD), yet the run times varied considerably depending on the parameter sets chosen. As an indication, the mean simulation run time for the first repeat in each catchment was 11 hours and 23 minutes for the Swale and 21 minutes for Tin Camp Creek. We used a batch mode functionality of CAESAR-Lisflood to run simulations of each repeat (16 model runs each) consecutively, and distributed batches across different machines – this is feasible for the model set ups described. However, for long-term simulations for catchments the size of the Upper Swale, individual model runs can take several weeks and running several runs consecutively becomes prohibitive. One solution would be to distribute the jobs on High Performance Computing (HPC) facilities, where the time for a single model run would not significantly decrease, but several hundred, even thousands, of individual model runs can be performed coincidently.

Here, t~~T~~he methodology has only been applied to the CAESAR-Lisflood model, and although some of the findings ~~will have implications to other LEMs, most~~ will be unique to CAESAR-Lisflood and the model set ups presented, they have implications for all LEMs. Importantly, t~~T~~he methodology ~~should~~

can serve as a highly useful tool for users to determine the behaviour of ~~each~~ any LEM model set up prior to calibration and/or simulation. ~~For CAESAR-Lisflood itself, future SA should analyse more catchments of different sizes and environmental conditions. The two model set ups used here should be analysed again but using a long term timeframe to understand how the model behaviour might evolve over longer simulations.~~

5. Conclusions

The feasibility of performing global SA to a highly parameterised catchment LEM has been demonstrated through the application of the MM to the CAESAR-Lisflood model. The ~~test~~ analysis was repeated over two different catchments suggesting some model behaviours are universal, and others vary depending on the catchment characteristics providing crucial information to inform future model developments. This analysis confirms that the sediment transport formulae are a significant source of uncertainty in LEMs, and in the CAESAR-Lisflood model the use of one formula over another can result in an order of magnitude differences in sediment yields when all other factors are kept constant. Another finding with relevance to SA and calibration of LEMs was the influence of parameters on each model function, showing that one aspect of model behaviour (e.g. catchment sediment yield) is not fully reflective of other, albeit related, model behaviours (e.g. internal geomorphology).

In addition to the above, the results reveal the parameters in CAESAR-Lisflood which exert the greatest influence, and whilst we can only apply this to the CAESAR-Lisflood model itself, it is likely that LEMs with comparable parameters will display similar behaviours. Some of the most influential parameters, like Manning's n roughness coefficient, grain size distributions, and vegetation critical shear stress are physically-based, so any uncertainty can be reduced by more detailed field measurements. We also show that parameters that determine the numerical efficiency of CAESAR-Lisflood exert a medium influence on the simulation results. Although some parameters exerted less influence on model behaviour relative to others, there were no parameters which did not influence the model in some way.

~~In addition to the above, the results reveal the parameters in CAESAR-Lisflood which exert the greatest influence, and whilst we can only apply this to the CAESAR-Lisflood model itself, it is likely that LEMs with comparable parameters will display similar behaviours. Some of the most influential parameters, like MNR, GSS and VEG are physically-based, so any uncertainty can be reduced by gathering data from the field – in these tests each of these parameters utilised global values initially, so more detailed field measurements could be utilised to provide spatially distributed values and further reduce uncertainty. The parameters which are most likely to be an issue for operators are those which have a medium influence and are set based on data characteristics for numerical efficiency – these include IOD, MinQ and MaxQ. For example, the typical and recommended value for MinQ is 1/100 of the DEM resolution and here, by varying the value yet keeping resolution the same, some variation was observed in the results – it is not yet determined whether any difference in model results at different resolutions are due to changes in values of MinQ and MaxQ, or the grid resolutions, or a combination of the two, and this will be a focus for future work.~~

The application of a global SA should become a vital step in any investigation using LEMs. This paper has demonstrated that the use of the MM is efficient for this purpose and yielded some ~~useful~~ valuable insights into model behaviour that can ~~be fed~~ultimately feed back into ~~the~~ model set up, ~~and~~ as well as future model development.

Model and Data Availability

The data produced by this study is made available on request from the corresponding author. The CAESAR-Lisflood model used in this study is freely available under a GNU licence from http://www.coulthard.org.uk

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors wish to thank the two reviewers, Andy Wickert and Daniel Hobley, for their insightful and helpful comments which have improved this manuscript. The Landscape Evolution Model Sensitivity Investigation Project (LEMSIP) ~~project~~ has emerged from the Field and Computer Simulation in Landscape Evolution (FACSIMILE) network. The aims of the project are to collate and generate knowledge pertaining to the sensitivities and uncertainties associated with Landscape Evolution Models, and how these influence the simulation of landscape development. The authors wish to thank the Young Geomorphologists group who donated computational resource. This work was supported by the NERC Flooding from Intense Rainfall (FFIR) project, Susceptibility of Basins to Intense Rainfall and Flooding (SINATRA) NE/K008668/1. The CAESAR-Lisflood model used in this study is freely available under a GNU licence from http://www.coulthard.org.uk

References

Adams, J. M., Gasparini, N. M., Hobley, D. E. J., Tucker, G. E., Hutton, E. W. H., Nudurupati, S. S. and Istanbulluoglu, E.: The Landlab v1.0 OverlandFlow component: a Python tool for computing shallow-water flow across watersheds, Geosci. Model Dev, 10, 1645–1663, doi:10.5194/gmd-10-1645-2017, 2017.

Andersen, J. L., Egholm, D. L., Knudsen, M. F., Jansen, J. D. and Nielsen, S. B.: The periglacial engine of mountain erosion - Part 1: Rates of frost cracking and frost creep, Earth Surf. Dyn., 3(4), 447–462, doi:10.5194/esurf-3-447-2015, 2015.

Armitage, J. J., Whittaker, A. C., Zakari, M. and Campforts, B.: Numerical modelling landscape and sediment flux response to precipitation rate change, Earth Surf. Dyn. Discuss., (May), 1–31, doi:10.5194/esurf-2017-34, 2017.

Aronica, G., Bates, P. D. and Horritt, M. S.: Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE, Hydrol. Process., 16(10), 2001–2016, doi:10.1002/hyp.398, 2002.

Attal, M., Tucker, G. E., Whittaker, A. C., Cowie, P. A. and Roberts, G. P.: Modelling fluvial incision and transient landscape evolution: Influence of dynamic Channel adjustment, J. Geophys. Res. Earth Surf., 113(3), 1–16, doi:10.1029/2007JF000893, 2008.

Di Baldassarre, G., Schumann, G. and Bates, P. D.: A technique for the calibration of hydraulic models using uncertain satellite observations of flood extent, J. Hydrol., 367(3), 276–282, doi:10.1016/j.jhydrol.2009.01.020, 2009.

Bates, P. D., Horritt, M. S., Aronica, G. and Beven, K.: Bayesian updating of flood inundation likelihoods conditioned on flood extent data, Hydrol. Process., 18(17), 3347–3370, doi:10.1002/hyp.1499, 2004.

Bates, P. D., Horritt, M. S. and Fewtrell, T. J.: A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling, J. Hydrol., 387(1–2), 33–45, doi:10.1016/j.jhydrol.2010.03.027, 2010.

Beven, K. and Kirkby, M.: A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d’appel variable de l’hydrologie du bassin versant, Hydrol. Sci. J., 24(1), 37–41 [online] Available from: http://www.tandfonline.com/doi/abs/10.1080/02626667909491834 (Accessed 8 May 2014), 1979.

Brockmann, D. and Morgenroth, E.: Comparing global sensitivity analysis for a biofilm model for two-step nitrification using the qualitative screening method of Morris or the quantitative variance-based Fourier Amplitude Sensitivity Test (FAST), Water Sci. Technol., 56(8), 85–93, doi:10.2166/wst.2007.600, 2007.

Campforts, B., Vanacker, V., Vanderborght, J., Baken, S., Smolders, E. and Govers, G.: Simulating the mobility of meteoric 10 Be in the landscape through a coupled soil-hillslope model (Be2D), Earth Planet. Sci. Lett., 439, 143–157, doi:10.1016/j.epsl.2016.01.017, 2016.

Campolongo, F., Cariboni, J. and Saltelli, A.: An effective screening design for sensitivity analysis of large models, Environ. Model. Softw., 22(10), 1509–1518, doi:10.1016/j.envsoft.2006.10.004, 2007.

Cariboni, J., Gatelli, D., Liska, R. and Saltelli, A.: The role of sensitivity analysis in ecological modelling, Ecol. Modell., 203(1–2), 167–182, doi:10.1016/j.ecolmodel.2005.10.045, 2007.

Collins, D. B. G.: Modeling the effects of vegetation-erosion coupling on landscape evolution, J. Geophys. Res., 109(F3), 1–11, doi:10.1029/2003JF000028, 2004.

Coulthard, T., Hicks, D. and Wiel, M. Van De: Cellular modelling of river catchments and reaches: Advantages, limitations and prospects, Geomorphology, 90(3–4), 192–207, doi:10.1016/j.geomorph.2006.10.030, 2007a.

Coulthard, T., Neal, J., Bates, P., Ramirez, J., de Almeida, G. and Hancock, G.: Integrating the LISFLOOD-FP 2D hydrodynamic model with the CAESAR model: implications for modelling landscape evolution, Earth Surf. …, 38(15), 1897–1906, doi:10.1002/esp.3478, 2013.

Coulthard, T. J. and Macklin, M. G.: How sensitive are river systems to climate and land-use changes? A model-based evaluation, J. Quat. Sci., 16(4), 347–351, doi:10.1002/jqs.604, 2001.

Coulthard, T. J. and Skinner, C. J.: The sensitivity of landscape evolution models to spatial and temporal rainfall resolution, Earth Surf. Dyn., 4(3), 757–771, doi:10.5194/esurf-4-757-2016, 2016a.

Coulthard, T. J. and Skinner, C. J.: The sensitivity of landscape evolution models to spatial and temporal rainfall resolution, Earth Surf. Dyn. Discuss., 1–28, doi:10.5194/esurf-2016-2, 2016b.

Coulthard, T. J. and Van De Wiel, M. J.: Quantifying fluvial non linearity and finding self organized criticality? Insights from simulations of river basin evolution, Geomorphology, 91(3–4), 216–235, doi:10.1016/j.geomorph.2007.04.011, 2007.

Coulthard, T. J. and Van De Wiel, M. J.: Modelling river history and evolution, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., 370(1966), 2123–2142, doi:10.1098/rsta.2011.0597, 2012.

Coulthard, T. J. and Van De Wiel, M. J.: Climate, tectonics or morphology: What signals can we see in drainage basin sediment yields?, Earth Surf. Dyn., 1(1), 13–27, doi:10.5194/esurf-1-13-2013, 2013.

Coulthard, T. J. and Van De Wiel, M. J.: Modelling long term basin scale sediment connectivity, driven by spatial land use changes, Geomorphology, 277, 265–281, doi:10.1016/j.geomorph.2016.05.027, 2017.

Coulthard, T. J., Lewin, J. and Macklin, M. G.: 12 Non-stationarity of basin scale sediment delivery in response to climate change, Dev. Earth Surf. Process., 11(07), 315–331, doi:10.1016/S0928-2025(07)11131-7, 2007b.

Coulthard, T. J., Ramirez, J., Fowler, H. J. and Glenis, V.: Using the UKCP09 probabilistic scenarios to model the amplified impact of climate change on drainage basin sediment yield, Hydrol. Earth Syst. Sci., 16(11), 4401–4416, doi:10.5194/hess-16-4401-2012, 2012.

Egholm, D. L., Andersen, J. L., Knudsen, M. F., Jansen, J. D. and Nielsen, S. B.: The periglacial engine of mountain erosion - Part 2: Modelling large-scale landscape evolution, Earth Surf. Dyn., 3(4), 463–482, doi:10.5194/esurf-3-463-2015, 2015.

Einstein, H. A.: The Bed-Load Function for Sediment Transportation in Open Channel Flows, Soil Conserv. Serv., (1026), 1–31 [online] Available from: https://ponce.sdsu.edu/einstein_bedload_function.pdf (Accessed 4 July 2018), 1950.

Fewtrell, T. J., Bates, P. D., Horritt, M. and Hunter, N. M.: Evaluating the effect of scale in flood inundation modelling in urban environments, Hydrol. Process., 22(26), 5107–5118, doi:10.1002/hyp.7148, 2008.

Fewtrell, T. J., Duncan, A., Sampson, C. C., Neal, J. C. and Bates, P. D.: Benchmarking urban flood models of varying complexity and scale using high resolution terrestrial LiDAR data, Phys. Chem. Earth, Parts A/B/C, 36(7), 281–291, doi:10.1016/j.pce.2010.12.011, 2011.

Gomez, B. and Church, M.: An Assessment of Bedload Sediment transport Formulae for Gravel Bed Rivers, Water Resour. Res., 25(6), 1161–1186, 1989.

Formatted: French (France)

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M. and Srinivasan, R.: A global sensitivity analysis tool for the parameters of multi-variable catchment models, J. Hydrol., 324(1–4), 10–23, doi:10.1016/j.jhydrol.2005.09.008, 2006.

Hall, J. W., Tarantola, S., Bates, P. D. and Horritt, M. S.: Distributed Sensitivity Analysis of Flood Inundation Model Calibration, J. Hydraul. Eng., 131(2), 117–126, doi:10.1061/(ASCE)0733-9429(2005)131:2(117), 2005.

Hancock, G. and Willgoose, G.: Use of a landscape simulator in the validation of the SIBERIA catchment evolution model: Declining equilibrium landforms, Water Resour. Res., 37(7), 1981–1992, doi:10.1029/2001WR900002, 2001.

Hancock, G. R.: The impact of different gridding methods on catchment geomorphology and soil erosion over long timescales using a landscape evolution model, Earth Surf. Process. Landforms, 31(8), 1035–1050, doi:10.1002/esp.1306, 2006.

Hancock, G. R. and Coulthard, T. J.: Channel movement and erosion response to rainfall variability in southeast Australia, Hydrol. Process., 26(5), 663–673, doi:10.1002/hyp.8166, 2012.

Hancock, G. R., Lowry, J. B. C., Coulthard, T. J., Evans, K. G. and Moliere, D. R.: A catchment scale evaluation of the SIBERIA and CAESAR landscape evolution models, Earth Surf. Process. Landforms, 35(8), 863–875, doi:10.1002/esp.1863, 2010.

Hancock, G. R., Coulthard, T. J., Martinez, C. and Kalma, J. D.: An evaluation of landscape evolution models to simulate decadal and centennial scale soil erosion in grassland catchments, J. Hydrol., 398(3–4), 171–183, doi:10.1016/j.jhydrol.2010.12.002, 2011.

Hancock, G. R., Lowry, J. B. C. and Coulthard, T. J.: Catchment reconstruction - erosional stability at millennial time scales using landscape evolution models, Geomorphology, 231, 15–27, doi:10.1016/j.geomorph.2014.10.034, 2015.

Hancock, G. R., Coulthard, T. J. and Lowry, J. B. C.: Predicting uncertainty in sediment transport and landscape evolution - the influence of initial surface conditions, Comput. Geosci., 90, 117–130, doi:10.1016/j.cageo.2015.08.014, 2016.

Hancock, G. R., Verdon-Kidd, D. and Lowry, J. B. C.: Sediment output from a post-mining catchment - Centennial impacts using stochastically generated rainfall, J. Hydrol., 544, 180–194, doi:10.1016/j.jhydrol.2016.11.027, 2017.

Herman, J. D., Kollat, J. B., Reed, P. M. and Wagener, T.: Technical Note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models, Hydrol. Earth Syst. Sci., 17(7), 2893–2903, doi:10.5194/hess-17-2893-2013, 2013.

Horritt, M., Bates, P. and Mattinson, M.: Effects of mesh resolution and topographic representation in 2D finite volume models of shallow water fluvial flow, J. Hydrol., 329(1–2), 306–314, doi:10.1016/j.jhydrol.2006.02.016, 2006.

Horritt, M. S. and Bates, P. D.: Effects of spatial resolution on a raster based model of flood flow, J. Hydrol., 253(1–4), 239–249, doi:10.1016/S0022-1694(01)00490-5, 2001.

Horritt, M. S. and Bates, P. D.: Evaluation of 1D and 2D numerical models for predicting river flood inundation, J. Hydrol., 268(1), 87–99, doi:10.1016/S0022-1694(02)00121-X, 2002.

Hunter, N. M., Horritt, M. S., Bates, P. D., Wilson, M. D. and Werner, M. G. F.: An adaptive time step solution for raster-based storage cell modelling of floodplain inundation, Adv. Water Resour., 28(9), 975–991, doi:10.1016/j.advwatres.2005.03.007, 2005.

Hunter, N. M., Bates, P. D., Neelz, S., Pender, G., Villanueva, I., Wright, N. G., Liang, D., Falconer, R.

A., Lin, B., Waller, S., Crossley, A. J. and Mason, D. C.: Benchmarking 2D hydraulic models for urban flooding, Proc. Inst. Civ. Eng. - Water Manag., 161(1), 13–30, doi:10.1680/wama.2008.161.1.13, 2008.

Ibbitt, R. P., Willgoose, G. R. and Duncan, M. J.: Channel network simulation models compared with data from the Ashley River, New Zealand, Water Resour. Res., 35(12), 3875–3890, doi:10.1029/1999WR900245, 1999.

Ijjasz-Vasquez, E. J., Bras, R. L. and Moglen, G. E.: Sensitivity of a basin evolution model to the nature of runoff production and to initial conditions, Water Resour. Res., 28(10), 2733–2741, doi:10.1029/92WR01561, 1992.

Istanbulluoglu, E. and Bras, R. L.: Vegetation-modulated landscape evolution: Effects of vegetation on landscape processes, drainage density, and topography, J. Geophys. Res. Earth Surf., 110(2), 1–19, doi:10.1029/2004JF000249, 2005.

Jerolmack, D. J. and Paola, C.: Shredding of environmental signals by sediment transport, Geophys. Res. Lett., 37(19), 1–5, doi:10.1029/2010GL044638, 2010.

Larsen, L., Thomas, C., Eppinga, M. and Coulthard, T.: Exploratory modeling: Extracting causality from complexity, Eos (Washington. DC)., 95(32), 285–286, doi:10.1002/2014EO320001, 2014.

Li, C., Zhang, L., Wang, H., Zhang, Y., Yu, F. and Yan, D.: The transferability of hydrological models under nonstationary climatic conditions., Hydrol. Earth …, 16(4), 1239–1254, doi:10.5194/hess-16-1239-2012, 2012.

Liu, B. and Coulthard, T. J.: Modelling the interaction of aeolian and fluvial processes with a combined cellular model of sand dunes and river systems, Comput. Geosci., 106, 1–9, doi:10.1016/j.cageo.2017.05.003, 2017.

Martin, Y. and Church, M.: Numerical modelling of landscape evolution: geomorphological perspectives, Prog. Phys. Geogr., 28(3), 317–339, doi:10.1191/0309133304pp412ra, 2004.

Met Office: 5km UK Composite Rainfall Data from the Met Office NIMROD System, NCAS Br. Atmos. Data Centre, available at : http://catalogue.ceda.ac.uk/uuid/82adec1f896af6169112d09cc1174499 (last access: 20 September 2016), 2003.

Morris, M. D.: Factorial Sampling Plans for Preliminary Computational Experiments, Technometrics, 33(2), 161–174, doi:10.2307/1269043, 1991.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I—A discussion of principles, J. Hydrol., 10, 282–290 [online] Available from: http://www.sciencedirect.com/science/article/pii/0022169470902556 (Accessed 8 May 2014), 1970.

Neal, J., Schumann, G., Fewtrell, T., Budimir, M., Bates, P. and Mason, D.: Evaluating a new LISFLOOD-FP formulation with data from the summer 2007 floods in Tewkesbury, UK, J. Flood Risk Manag., 4(2), 88–95, doi:10.1111/j.1753-318X.2011.01093.x, 2011.

Neelz, S. & Pender, G.: Benchmarking the latest generation of 2D hydraulic modelling packages. [online] Available from: http://evidence.environment-agency.gov.uk/FCERM/Libraries/FCERM_Project_Documents/SC120002_Benchmarking_2D_hydraulic_models_Report.sflb.ashx, 2013.

Neumann, M. B.: Comparison of sensitivity analysis methods for pollutant degradation modelling: A case study from drinking water treatment, Sci. Total Environ., 433(October), 530–537, doi:10.1016/j.scitotenv.2012.06.026, 2012.

Norton, J. P.: Algebraic sensitivity analysis of environmental models, Environ. Model. Softw., 23, 963–972, doi:10.1016/j.envsoft.2007.11.007, 2008.

Norton, J. P.: Selection of Morris trajectories for initial sensitivity analysis, IFAC., 2009.

Oakley, J. E. and O'Hagan, A.: Probabilistic Sensitivity Analysis of Complex Models : A Bayesian Approach Author (s): Jeremy E . Oakley and Anthony O'Hagan Published by : Wiley for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/3647504 Probabilistic sensitiv, , 66(3), 751–769, 2004.

Oreskes, N., Shrader-Frechette, K. and Belitz, K.: Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, Science (80-.)., 263, 641–646, doi:10.2307/2883078, 1994.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J. and Roo, A. P. J. De: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), Hydrol. Earth Syst. Sci. Discuss., 9(4), 381–393, doi:10.5194/hess-9-381-2005, 2005.

Pappenberger, F., Harvey, H., Beven, K., Hall, J. and Meadowcroft, I.: Decision tree for choosing an uncertainty analysis methodology : a wiki experiment, Hydrol. Process., 20, 3793–3798, doi:10.1002/hyp, 2006.

Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R. and Matgen, P.: Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations, Hydrol. Earth Syst. Sci. Discuss., 11(2), 739–752 [online] Available from: https://hal.archives-ouvertes.fr/hal-00305049/ (Accessed 24 May 2017), 2007.

Pappenberger, F., Beven, K. J., Ratto, M. and Matgen, P.: Multi-method global sensitivity analysis of flood inundation models, Adv. Water Resour., 31(1), 1–14, doi:10.1016/j.advwatres.2007.04.009, 2008.

Pazzaglia, F. J.: Landscape evolution models, pp. 247–274., 2003.

Petersen, A. C. (Arthur C.: Simulating nature : a philosophical study of computer-simulation uncertainties and their role in climate science and policy advice, CRC Press. [online] Available from: https://books.google.co.uk/books?hl=en&lr=&id=I4GhNkiPv3EC&oi=fnd&pg=PP1&dq=Simulating+Nature:+A+Philosophical+Study+of+Computer-Simulation+Uncertainties+and+Their+Role+in+Climate+Science+and+Policy+Advice&ots=EKmUbPTtVZ&sig=BisleTDNw3E0_EpozyLbxjJHUdg#v=onepage&q=Simulating Nature%3A A Philosophical Study of Computer-Simulation Uncertainties and Their Role in Climate Science and Policy Advice&f=false (Accessed 18 August 2017), 2012.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B. and Wagener, T.: Sensitivity analysis of environmental models: A systematic review with practical workflow, Environ. Model. Softw., 79, 214–232, doi:10.1016/j.envsoft.2016.02.008, 2016.

Pujol, G.: R Package "sensitivity". Version 1.4-0, 2009.

R Hancock, G.: Modelling stream sediment concentration: An assessment of enhanced rainfall and storm frequency, J. Hydrol., 430–431, 1–12, doi:10.1016/j.jhydrol.2012.01.022, 2012.

Ratto, M., Pagano, A. and Young, P.: State Dependent Parameter metamodelling and sensitivity analysis, Comput. Phys. Commun., 177(11), 863–876, doi:10.1016/j.cpc.2007.07.011, 2007.

Saltelli, A., Chan, K. and Scott, E. M.: Sensitivity Analysis, John Wiley, New York, 2000.

Sampson, C. C., Fewtrell, T. J., Duncan, A., Shaad, K., Horritt, M. S. and Bates, P. D.: Use of terrestrial laser scanning data to drive decimetric resolution urban inundation models, Adv. Water Resour., 41, 1–17, doi:10.1016/j.advwatres.2012.02.010, 2012.

Saynor, M. J., Lowry, J., Erskine, W. D., Coulthard, T. and Hancock, G.: Assessing Erosion and Run-Off Performance of a Trial Rehabilitated, Proc. Life Mine Conf. July 2012, (July), 10–12, 2012.

Skinner, C. and Coulthard, T.: Caesar-Lisflood Existing Applications Parameter Listings - May 2017, , doi:10.5281/ZENODO.800558, 2017.

Sobol', I.: Global Sensitivity Indices for Nonlinear Mathematical Models:Review, Math. Comput. Simul., 55, 271–280, doi:10.1016/S0378-4754(00)00270-6, 2001.

Song, X., Zhan, C., Xia, J. and Kong, F.: An efficient global sensitivity analysis approach for distributed hydrological model, J. Geogr. Sci., 22(2), 209–222, doi:10.1007/s11442-012-0922-5, 2012.

Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M. and Xu, C.: Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications, J. Hydrol., 523(225), 739–757, doi:10.1016/j.jhydrol.2015.02.013, 2015.

Stephens, E. M., Bates, P. D., Freer, J. E. and Mason, D. C.: The impact of uncertainty in satellite data on the assessment of flood inundation models, J. Hydrol., 414–415, 162–173, doi:10.1016/j.jhydrol.2011.10.040, 2012.

Tucker, G. E. and Bras, R. L.: A stochastic approach to modelling the role of rainfall variability in drainage basin evolution, Water Resour. Res., 36(7), 1953, doi:10.1029/2000WR900065, 2000.

Tucker, G. E. and Hancock, G. R.: Modelling landscape evolution, Earth Surf. Process. Landforms, 35(1), 28–50, doi:10.1002/esp.1952, 2010.

Vanwalleghem, T., Stockmann, U., Minasny, B. and McBratney, A. B.: A quantitative model for integrating landscape evolution and soil formation, J. Geophys. Res. Earth Surf., 118(2), 331–347, doi:10.1029/2011JF002296, 2013.

Welivitiya, W. D. D. P., Willgoose, G. R., Hancock, G. R. and Cohen, S.: Exploring the sensitivity on a soil area-slope-grading relationship to changes in process parameters using a pedogenesis model, Earth Surf. Dyn., 4(3), 607–625, doi:10.5194/esurf-4-607-2016, 2016.

Van De Wiel, M. J. and Coulthard, T. J.: Self-organized criticality in river basins: Challenging sedimentary records of environmental change, Geology, 38(1), 87–90, doi:10.1130/G30490.1, 2010.

Van De Wiel, M. J., Coulthard, T. J., Macklin, M. G. and Lewin, J.: Embedding reach-scale fluvial dynamics within the CAESAR cellular automaton landscape evolution model, Geomorphology, 90(3–4), 283–301, doi:10.1016/j.geomorph.2006.10.024, 2007.

Van De Wiel, M. J., Coulthard, T. J., Macklin, M. G. and Lewin, J.: Modelling the response of river systems to environmental change: Progress, problems and prospects for palaeo-environmental reconstructions, Earth-Science Rev., 104(1–3), 167–185, doi:10.1016/j.earscirev.2010.10.004, 2011.

Wilcock, P. R. and Crowe, J. C.: Surface-based Transport Model for Mixed-Size Sediment, J. Hydraul. Eng., 129(2), 120–128, doi:10.1061/(ASCE)0733-9429(2003)129:2(120), 2003.

Willgoose, G. R., Hancock, G. R. and Kuczera, G.: A Framework for the Quantitative Testing of Landform Evolution Models, pp. 195–216, American Geophysical Union., 2003.

Wong, J. S., Freer, J. E., Bates, P. D., Sear, D. A. and Stephens, E. M.: Sensitivity of a hydraulic model to channel erosion uncertainty during extreme flooding, Hydrol. Process., 29(2), 261–279, doi:10.1002/hyp.10148, 2015.

Yang, J.: Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis, Environ. Model. Softw., 26(4), 444–457, doi:10.1016/j.envsoft.2010.10.007, 2011.

Ziliani, L., Surian, N., Coulthard, T. J. and Tarantola, S.: Reduced-complexity modeling of braided rivers: Assessing model performance by sensitivity analysis, calibration, and validation, J. Geophys. Res. Earth Surf., 118(4), 2243–2262, doi:10.1002/jgrf.20154, 2013.